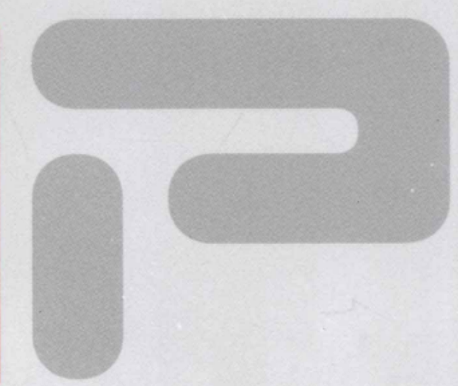
万卷方法 | 心理学研究方法丛书
中国心理学会心理学教学工作委员会推荐

INTRODUCTION TO RESEARCH METHODS IN PSYCHOLOGY



心理学研究方法导论

休·库利坎(Hugh Coolican) 著
卢家楣 等 译





重庆大学出版社
http://www.cqup.com.cn

YINLIWUE YANJIU FANGFA DAOLUN

本书作者运用轻松语气和妙趣横生的生活常识、研究实例，将学生对研究方法的恐惧一扫而光。

对任何一个心理学学生，学会如何做研究、如何认识设计的优缺点、如何避免不被骗人的结果和错误的统计方法所蒙蔽，是非常重要的。学习研究方法带给你的是技能，而不只是知识，并且这些技能在心理学课程之外也会令你受益。

科学的逻辑和推理，不是只有受过良好教育的人们才能使用的技能（否则就太悲哀了）。我们都拥有这种技能。在你很小的时候，为了理解和预测周围复杂多变的世界，你就一直在使用这种技能。……在教授研究方法甚至统计学的时候，我始终坚信，读者对本书讨论的许多观点都已形成了知觉概念。

——作者

在这本书中，作者借助常识、逻辑和日常生活经验，向学生阐明理解和成功进行研究的技能和技术。这种写作风格对于之前没有研究方法和统计经验的本科生和研究生，以及对这类学习带有恐惧心理的人来说颇有裨益。

参阅及发表相关评论请登陆“万卷方法与学术规范博客圈”
http://q.blog.sina.com.cn/fafang

上架建议：学术社科

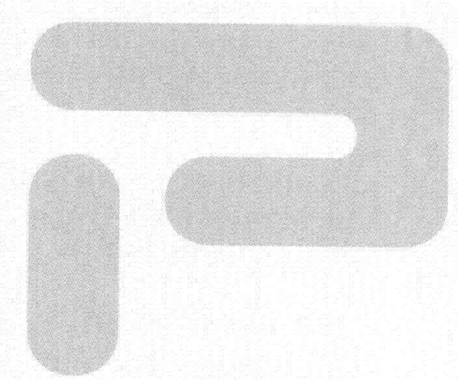
ISBN 978-7-5624-5828-9



定价: 39.80元

万卷方法 | 心理学研究方法丛书
中国心理学会心理学教学工作委员会推荐

INTRODUCTION TO
RESEARCH
METHODS
IN PSYCHOLOGY



心理学研究方法导论

休·库利坎(Hugh Coolican) 著
卢家楣 等 译

重庆大学出版社

图书在版编目(CIP)数据

心理学研究方法导论/(英)休·库利坎(Coolican,H.)著;卢家楣等译.—重庆:重庆大学出版社,2011.10
(万卷方法)
书名原文:Introduction to Research Methods in Psychology
ISBN 978-7-5624-5828-9

Ⅰ.①心… Ⅱ.①库…②卢… Ⅲ.①心理学研究方法 Ⅳ.①B841

中国版本图书馆 CIP 数据核字(2010)第 233830 号

心理学研究方法导论
(英文第 3 版)
[英]休·库利坎 著
卢家楣 等 译
策划编辑:林佳木
责任编辑:杨 敬 邬小梅 版式设计:雷少波
责任校对:邹 忌 责任印制:赵 晟
*
重庆大学出版社出版发行
出版人:邓晓益
社址:重庆市沙坪坝区虎溪大学城重庆大学(虎溪校区)
邮编:401331
电话:(023)88617183 88617185(中小学)
传真:(023)88617186 88617166
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn(营销中心)
全国新华书店经销
重庆升光电力印务有限公司印刷
*
开本:787×1092 1/16 印张:17.25 字数:377千 插页:16开2页
2011年10月第1版 2011年10月第1次印刷
印数:1—4 000
ISBN 978-7-5624-5828-9 定价:39.80元

作译者简介

休·库利坎(Hugh Coolican)　英国考文垂大学(Coventry University)心理学主讲师,注册心理学家和国际学士学位评定的助理主考官。他是海德兰(Hodder Arnold)出版社出版的畅销书《心理学研究方法和统计》第四版的作者。

卢家楣　上海师范大学教授,国家级教学名师,原教育科学学院院长,现上海师范大学心理研究所所长,博士点学科带头人,博士后流动站站长,心理学重点实验室主任,中国心理学会常务理事,中国心理学会教育心理学分会副会长,中国心理学会教学工作委员会副主任,全国教育科学规划心理学学科评审组成员,全国教育专业学位教育指导委员会委员,全国中小学心理健康教育专家指导委员会委员,《心理科学》副主编,主要从事教育心理学、青少年心理学、普通心理学的教学和研究,并开拓了情感教学心理学领域,提出"以情优教"的教学思想、教学原则、教学模式、教学策略和教学评价,发表著作12部,论文150多篇,曾获全国和省部级科研和教学优秀成果奖20余项。

总 序

自古以来,人类一直在探索着自己的内心世界:我到底是个什么样的人?为什么在许多方面我与周围人们如此相似,而在其他方面又如此不同?我们是怎样认知世界的?为什么有时记忆会错误有时记忆会很牢固?喜怒哀乐、爱恋、责任感是如何产生的?情绪能自我调控吗?意识是怎么回事?梦是怎么回事,它能预测未来吗?为什么一个人独处时和在群体中的行为是不一样的?我们是怎样理解语言的,又是怎样组织和表达语言的?如此等等的问题,从传说、神话、甲骨文中我们都能发现,可见,人类长期以来一直对理解自身的内心世界有着浓厚的兴趣。然而直至1879年冯特(Wilhelm Wundt, 1832—1920)在莱比锡大学建立了心理实验室之后,人类才开始借助科学方法来寻求这些问题的答案。通过精密严格的数据收集与事实分析来研究心理与行为,积累知识,从而发展出今日的心理学。而随着心理学的发展,其研究方法也发展起来。

心理学家的科学研究是一种自觉的、有目的地探索精神世界的求知活动。这种探索,不仅有理论,还有与理论有关的观点及方法、仪器等。当代心理学主要有五种理论取向在探讨人类内心世界的奥秘,并对大脑如何工作提出了不同的假设,因而它们所采用的方法也不同。持生物学理论取向的心理学家认为心理是脑的机能,采用脑电图(EEG)、正电子发射断层扫描技术(PET)、功能核磁共振成像技术(FMRI)等方法来探讨人类是如何产生知觉、记忆、推理、情绪和某些人格特征的。持学习理论取向的心理学家认为心理是个体对环境条件作用积累经验的生理变化,采用操作条件作用、奖赏、惩罚、观察学习等方法来探讨人类和动物行为的形成及矫正。持认知理论取向的心理学家把人类的心理活动视为类似于计算机的信息加工,用反应时、正误率和口语报告等手段来探讨人类的知觉、记忆、言语和思维等心理过程。持精神分析论取向的心理学家把人的心理视为潜意识本能的表现,采用个案调查、诠释学方法来探讨个人内部的驱力、冲突或心理疾病等潜意识活动。持人本主义理论取向的心理学家把人的心理视为自我实现需求的表现,采用相关法、诠释学方法来探讨自由意志、个人成长、潜能实现等问题。在心理学研究中理论观点与方法始终是结合在一起,相辅相成的。它们既是指导这种探索活动的武器,又是保证这种活动取得成果的基础。正因为有了这一套套理论观点与方法的有机结合,心理学的科学研究才成为一种自觉的、有目的的定向活动,心理学也才成其为科学。因此,学习心理学研究方法具有十分重要的意义。

首先,有助于我们自觉地将理论与方法相结合,养成科学思维的习惯。心理学家的科学研究都具有明确的目的,即需要解决的问题,为此,就要对该问题以往的研究和目前的现状进行文献综述,并按照一定的有效程序对其进行探讨。即是说,心

理学研究的基本程序与任何科学研究是一样的，都包含下列步骤：选题和提出假设——设计研究方案（用以检验假设的真伪）——收集资料——整理和分析资料——解释结果和检验假设。从心理学研究程序的各个环节可以看出，在心理学研究中，精密的仪器和先进的实验设备固然重要，然而最重要的还是研究者的头脑。通过对心理学研究方法的学习，将有助于我们养成善思考和科学思维的习惯。心理学研究方法，不仅可以帮助我们运用自己的智慧去进行科学研究，而且它还可以帮助我们去鉴别自己和他人的研究成果的正确与谬误。正因为如此，有经验的学者在评价一篇学术论文时往往不只是看它的结论，而且，甚至是更重要的还要看论文作者是通过怎样的途径和方法而获得结论的。

其次，有助于激发我们的创新观念和达成创新目标。心理学的理论、观点和研究方法是多样性的，其研究成果的科学性也不同。学习了心理学研究方法之后，我们了解到心理学研究的各种方法，如个案法、相关法和实验法在心理学研究的不同时期有不同的用途，其信效度也是不同的。个案研究在心理学研究初期是有用的（有助于发现可供研究的现象和变量），但是要确定变量之间的因果关系，建立科学理论，则必须借助于实验法。弗洛伊德（Sigmund Freud, 1856—1939）根据自己对歇斯底里病人的临床观察和对梦、失误和笑话等的现象分析，建立起以潜意识动机为基础的精神分析理论。这个理论不是一个科学的理论，其中许多概念和命题缺乏实证效度。然而，熟悉实验法的学者想到用实验法来检验弗氏理论中的许多概念。例如，对于潜意识这种现象，他们发展出一系列的内隐实验程序来进行检验，结果发现有内隐学习、内隐记忆、内隐情绪、内隐动机的存在。虽然20世纪80年代兴起的心理潜意识研究与弗洛伊德的潜意识的性本能和死亡本能有本质的区别，但却加深了我们对心理成分和潜意识性质的认识。自我是心理学研究的一个主题，在科学心理学的早期，一些著作从理论上探讨自我的性质，到了20世纪70年代不少学者开始用相关法和实验法探讨自我的成分和机能，后来由于引入神经科学方法（如脑的电刺激、功能核磁共振成像技术）才开始探索自我的脑机制。正如巴甫洛夫（И. П. Павлов, 1849—1936）所说“科学是随着研究方法所获得的成就而前进的。研究方法每前进一步，我们就更提高一步。随之在我们面前也就开拓了一个充满着种种新的、更加广阔的远景。因此，我们头等重要的任务乃是制定研究方法。”①

第三，有助于年轻心理学工作者快速成长。做任何事情都要讲究方法。方法对头，事半功倍；方法不对，事倍功半，甚至导致失败。心理学史表明，有些心理学家之所以能在学科上有所建树、有所贡献，除了他们的天赋聪慧，当时的科技水平和良好的学术环境外，还往往与他们能正确运用新的研究方法有密切的关系。系统地学习心理学研究方法显然比只凭个人经验、漫无边际地去摸索，更能促进年轻心理学工作者的快速成长。

2005年秋天，重庆大学出版社“万卷方法”总策划雷少波同志带着已出版的新书来征求我对翻译这套图书的意见。我很高兴看到他们对这套书的设想：“万卷方法”是重庆大学出版社从2004年开始出版，拟系统深入地介绍各门社会科学研究方法的大型工具性丛书，其中包括心理学研究方法。这是一项促进我国心理学事业发

---

① 巴甫洛夫选集，北京：科学出版社，1955：49

展的开创性工作，我给予热情鼓励和支持。在我看来，“万卷方法——心理学研究方法丛书”具有以下特点：

（1）品位高。对于研究方法的著作来说，质量优、品位高是最重要的。丛书所介绍的是国外心理学领域中，许多有成就的心理学家所普遍认可的心理学研究工作的原理、方法，研究课题设计，以及如何正确使用各种可以应用的技术手段等。例如如何做心理学实验、如何进行心理学的质性研究、如何撰写心理学学术论文，以及在心理研究中如何使用数理统计、应遵循哪些伦理道德等。其中像心理学质性研究、心理学研究伦理道德等著作，国内至今未见有专题著作翻译出版，是国内急需的。

（2）适用面广。丛书所介绍的心理学研究方法是相对基础性的，可供高校心理学专业的本科生和研究生作为教材或教学参考书使用，也可供广大人文社会科学工作者参考。

（3）开放性。根据我国心理学教学和科研的需要以及心理学研究方法的发展，出版社将通过版权引进和本土开发，使丛书不断丰富与完善。

我相信广大读者会喜爱“万卷方法——心理学研究方法丛书”，祝愿“万卷方法”不断发展，日益完善。

是为序。

黄希庭

谨识于2009年10月

西南大学 有容斋

译者前言

古人云“工欲善其事，必先利其器”。的确，工具质量的优劣会决定事倍功半还是事半功倍。对于研究人员来说，研究方法的水平同样会影响研究工作的成效。心理学研究也不例外，甚至可以说，心理学研究在某种程度上更依赖于研究方法的科学性水平。这是因为心理学研究的对象——人的心理现象，是宇宙间最神奇的现象之一，具有高度的复杂性。当初正是德国心理学家冯特在莱比锡大学创建了世界上第一个心理实验室，运用了自然科学的方法来研究人类的心理现象，实现了心理学研究方法上的革新，才使心理学摆脱了哲学的藩篱，走上了独立学科的发展征程，掀开了科学发展的新篇章。嗣后，心理科学始终伴随着研究方法与技术的创新而演进。

近年来，心理学在我国的进展很快，师范大学大多建立了心理学教学和研究机构，不少综合性大学也设立了心理系，有的大学还成立了心理学院，使从事心理学学习的人数越来越多，研究队伍也越来越壮大，对心理学研究方法方面的知识的需求也越来越强烈。然而心理学出版物虽日见增多，心理学研究方法类的书籍却十分匮乏。据有人不完全统计，我国心理学工作者撰写的以心理学研究方法为题的书籍约10种左右，加上影印的外文版及译著也不超过20种。这些书籍为推动心理学研究的高水平、科学化发展发挥了极大的作用，但是在书籍的品种和数量上无法满足客观实际的需要。另一方面，由于研究方法一类书籍本身的特点，加上作者严谨慎密的写作风格，往往易使学习者感到枯燥、艰涩，特别是对初学者来说，更易产生望而生畏之感。因此，迫切需要有一些通俗易懂、操作实用的心理学研究方法类的书籍问世，以飨读者。正是在这样的背景下，英国学者休・库利坎撰写的《心理学研究方法导论》进入我们的视线，顿有一阵春风拂面的清馨之感。

作者休・库利坎是英国考文垂大学(Coventry University)心理学主讲师，注册心理学家和国际学士学位评定的助理主考官。他长期从事心理学研究方法的教学和研究工作，有着这方面丰富的授课和写作经验，也是海德兰(Hodder Arnold)出版社出版的畅销书《心理学研究方法和统计》第四版的作者。他谙熟初学者学习过程中易遇到的问题和出现的心态，又精通心理学研究方法，因此从内容的选择到内容的呈现，从论述的方式到论述的语句，从理论阐述到实例应用，处处体现了对学习者的理解和对研究方法的彻悟。因此，本书是一部集科学性、实用性、丰富性、可读性、趣味性于一体的著作。当我们第一次看到原版时，就被吸引住了。在翻译过程中，我们进一步体会到本书独特的魅力，感受到先读为快的愉悦。难怪本书在英国甚至西

方有较大的影响，被选为许多大学的教科书，在读者中享有盛誉。具体说，本书有以下几个特色：

1. 严谨而不失生动

本书作为一本学术著作，处处体现着严谨与规范，从概念的解释到文献的引用，从体例的安排到文字的表述，都给我们一种严谨、科学的感受。因此，阅读本书首先将会得到一种科学的熏陶。然而，与一般介绍方法的书籍不同的是，本书中有大量的案例与专栏，这些案例和专栏又与人们日常生活紧密联系，帮助学习者更好地理解相关内容。同时，在行文中，在案例描述中作者以其特有的诙谐幽默的笔触和生动活泼的语气，一扫枯燥、呆板之感，给人一种轻松、愉悦的体验。

2. 全面而不失重点

本书可以看作是一本心理学研究方法的入门指导书。从“心理学与研究”作为总论开篇，到“实践研究的计划和研究报告的撰写”作为全书收篇，内含研究变量和样本确定、实验和准实验的方法比较、观察法和访谈法实施、问卷和量表编制、定量和定性研究、数据收集和处理、显著性和差异性检验、各种相关分析、心理学研究中道德问题的关注等，作为心理研究的基本内容，已十分全面，给学习者一个完整的心理学研究方法的框架体系。但在全面介绍心理学研究方法时，本书又非泛泛而谈、平铺直叙，而是突出重点、难点与盲点。从整体篇幅来看，一是在介绍定量研究与定性研究的同时，突出定量研究，因为定量方法仍然是心理学研究的主流方法；二是在整体介绍的同时，在第一章和第二章中突出对心理学研究的背景与要义的介绍，使得读者不仅了解“其然”，而且了解“其所以然”；三是强化心理学研究中的道德理念，诸如研究中对被试的尊重，研究结果的公开等都会涉及伦理道德问题，并在法律完善及人权意识增强的趋势下，更凸显其重要性。总之，这样点面结合的处理方式，使本书充分体现了书名所冠以的“导论”之涵义。

3. 理性而不失直观

本书对于方法论及各种方法的特点及其应用，均有科学的分析，包括作者自己的观点，体现了科学思维与逻辑推演，从而透射出理性的光芒。但是，在理性论述的同时，本书为我们呈现了大量的插图与表格。这些插图与表格不仅使我们更容易理解有关概念，而且学习到如何直观说明问题的诀窍。如对于样本、样本分配、有偏样本、负相关、相关度等概念的论述都有十分直观的图示。即使没有任何统计学背景的读者，也能够一目了然。又如，尽管数据分析与统计检验在大部分心理学方法书籍里都有介绍，但本书直观的图示、通俗的表述、翔实的案例对学习者更富吸引力。同时，本书在重要概念上均用不同字体和加粗方式显示，以引起读者的注意，符合心理学原理。

4. 宏论而不失操作

本书的介绍从宏观入手，首先谈了心理学研究的意义、思路、方向和理念，然后介绍了心理学研究方法的整体框架体系，给人以一种统摄全局、高屋建瓴之感。但是，作为一种研究方法的书籍，本书充分体现了操作性和实用性的特点。在论述诸

如样本的生成、观察法的注意点、问卷的编制、显著性检验的选择、研究计划及研究报告的撰写等内容时，作者几乎是在手把手地教读者入门。这种细致入微的富有实用性和操作性的解释，极利于学习者在学习时理论不觉抽象，在运用中"拷贝不会走样"。

5. 实训而不忘鼓励

本书作为心理学本科生、研究生以及其他初学者的学习教材，也可作为心理学专业研究者的参考书籍，非常注意学习者的实际训练。每章一开始均有重点内容提示，便于读者从总体把握本章内容。每章节后面全部附有练习，而且均有参考答案，这便于学习者对所学内容进行自我检测，为学习内容的理解与巩固提供了条件。在章节结束时还有关键术语，使学习者再次对本章的框架体系及重要概念进行回顾。还要提到的是，作者针对初学者在学习心理学研究方法时易产生畏难、退缩情绪的实际情况，在全书撰写中非常注意为读者传递这样的信息：即便在之前没有研究方法和数理统计方面基础的学生，凭借生活常识、逻辑思维和日常经验，也已具备了理解内容并成功进行研究的技能和技术。这就为学习者树立学习上的自信。真不愧心理学专家撰写的心理学教材！

本书虽然是心理学研究方法类的教材，但对教育科学研究者也是适用的。因为现今的教育科研也逐渐强调定性研究和定量研究的结合，许多心理学研究的方法也完全适用于教育科学研究。当前，我国教育科学研究方法的书出了不少，但仍不能满足实际的需求与方法进展的需要。前一阶段，重庆大学出版社出版了万卷方法丛书，受到社会科学界的赞誉及读者的好评。而出版者在万卷方法策划报告暨出版说明中所写的"为方法理性鼓与呼"，也正反映了他们对中国社会科学界研究方法类书籍缺乏的担忧，反映了他们对中国社会科学进一步发展的支持。中国社科院沈崇麟先生在社会科学研究方法经典译丛所写的总序中也谈到，相对于国内蓬勃开展的社会科学研究及国外迅猛发展的研究方法而言，我国社会科学研究方法的进展是滞后的。这就需要进行深入的科学研究。同样，这本书还可作为教育实践工作者、中小学教师进修的读物。当前，随着我国教育的快速发展和基础教育改革的不断深入，对中小学教师专业化发展的要求也越来越高，其中一个重要的方面就是要提高教育科研能力，要成为研究型的教师。这样本书所提供的心理学研究理论和技术完全可以适用于教育研究，可用来研究课堂教学中有效教学的方法及其测评手段，研究学生的学习心理、品德心理、健康心理、差异心理、学习困难等问题，从而为提高教书育人的科学性和艺术性，为促进学生素质的全面而有个性化的发展作出积极的贡献。

鉴于本书上述特色，我欣然接受中国心理学会心理学教学工作委员会的推荐，承担本书的翻译工作，并作为上海市重点学科"发展与教育心理学"建设项目（S30401）的一个任务来加以完成。我的博士生巫文胜、张敏、张文海、王俊山、田学英以及硕士生李志专、嵇家俊、崔毓婕、李玲玲、张燕燕、徐京卫、周栋梁等竭尽全力帮助我完成翻译和初校工作，其中巫文胜、李玲玲还协助我进行必要的组织工作。

此外，我的学生许洋、孙俊才和张奇勇先后参与我的审阅和校对工作。可以说，没有他们的积极、认真、高效的努力，此书是无法在近期完成的。在付梓之际，还要感谢重庆大学出版社领导和雷少波编辑在整个译著的翻译、出版过程中所给予的指导和帮助，感谢本书作者库利坎先生写出这么精彩的书稿，使我们在翻译的过程中乐此不疲，且受益良多。

但愿此书不仅对心理学专业的学生、研究人员，而且对教育科研工作者和有志于提高教育科研水平的中小学教师都有所裨益，我们将会因此感到由衷的欣慰。诚然，在翻译中不免有词不达意、言不达雅之处，还望读者原宥、赐教。

卢家楣
上海师范大学
2011 年 8 月

英文版前言

嗨，让我先来猜一下。你之所以没有翻开这本书，是因为你在寻找一本能让你爱不释手，可以消磨在海滩或游船度假时光的好书，抑或是一本振奋人心、拓展思维的好读物。虽然我并不希望这样，但学生们往往不会满怀热情地参加研究方法的讲座。事实上，他们对待讲座的态度和我把蛋奶冻的外皮留到最后再吃是一样的。学生们通常对心理学家们所研究的结果很着迷，然而心理学家们是**如何**获得这些发现的呢，这就似乎难以引起同样的兴趣了。

对上述现象，我感到很遗憾。首先，只有当学生们学会自信满满地提问"心理学家们是如何知道的?"时，教育才能被视为是成功的。其次，一门好的研究方法课程应该传授给学生的不仅仅是知识，更是研究的技能。当下的你正值青春年华，我可以向你保证这些技能在你今后的事业和生活中定有用武之地。你们中的很多人被要求去发现别人在想些什么，而你做这事的方式一定是心理学的研究方法。

事实上，除了统计数据外，大多数研究方法的问题和概念是关于人们在研究情境中是如何反应的：我们如何才能获取最好的信息？人们在被观察时是如何作出反应的？人们在测试时是否会努力表现得更好些？诸如此类。因此，和许多在你的心理学教学大纲里看上去较有趣味性的领域一样，各种研究方法也会使你专注于对人们是如何思维和行为问题的思考，而你也就在进行心理学研究了。

让我们试想一下，为何你会翻看这本书。我猜想可能是因为你们中的大多数人已经在学习 AS/A 级别的心理学课程，抑或是学位课程中包含心理学内容，也可能是你正在学习一门人际关系的课程，而课程要求你去思考该如何收集他人的信息。你们的导师可能提供了一张阅读清单或者你已经作了相关的研究，并发现这本书对你的研究颇有裨益。如果是这样的话，欢迎你加入心理学之旅！通过对方法和数据的讲解，我的学习指引始终力图使你的学习之旅轻松有趣，甚至那一个个清晰明确、源自生活的具体实例也会有助于你自学成长。经过多年的教学，我依然深信当学生们来听心理学课时，他们头脑中已经形成了一些理念，并期望从方法中获得进一步的汲取。这些理念之所以已形成，是因为学生们有来自生活和待人接物的经验。许多科幻小说只是在生活常识的基础上，加上了对人类及其行为的更严格、细致的探索。因此，我希望这本书能帮助你们亲自探究这些理念，并能生动地描绘它们，以冀让你们更为充分地了解人类研究的全貌。

本书更进一步的强烈设想就是能使数据的统计处理变得易于学习和入门。许多学生仍然认为他们不太有数学细胞，但却惊喜地发现他们可以仿照、操作，甚至理解书中数据统计处理的过程，而这一过程所需要的数学运算，仅仅只停留于加、减、乘、除的要求。这要求也的确就是你所需要掌握的全部。

本书作为第三版，已打破了章节间的限制，让你能在对一大块信息的分步消化吸收后，用一套练习加以测试。那些进步反馈框可能会在一个章节中出现3至4次，并要求你回答问题，关注重点术语，而不是让你一下子消化吸收整个章节。在书后的词汇表中都列有重点术语的定义。这个词汇表会告诉你，在AS/A2级别的心理学中，哪些重点术语会具体出现在4个教学大纲中的哪一个，以便让你精确地了解在参加A1或A2水平考试前，你应该正确掌握哪些术语。每一个章节都始于要点总结，这让你知道你正在学习的内容或提醒你，一旦你已学习了该章节，应该掌握的内容是什么，以及在学习的过程中的基本要素有哪些。

许多学生会有这样的感受，针对实际的观察写一份科学报告是他们目前在所受教育中需要面对的最为恐怖的任务之一。所有的A水平课程要求至少一份报告，而大多数学位课程在第一年要求几份报告。为了能支持和帮助你完成这样的任务，本书中有一个完整章节是关于如何计划和撰写报告的，这其中融入了我在多年教学中了解到的学生常犯的错误、会遇到的陷阱、应包括或不应包括的内容等大量的知识。然而，没什么比实例更清晰明了的了，为此书中提供了不是一份而是两份报告——一份“一般”报告附有想象记号(实际也是如此)的完整评语；一份来自同一实践工作的“好”报告。遵照书中这部分的要求，且运用较好的语言来表达，你定会取得不错的成绩。

最后，大多数A水平课程和学位课程包括一些书面形式的测评，其形式涵盖了通过阅读实际研究内容中的一段场景，并就实验项目的几个问题给出简短回答。本书为你准备了7道类似的问题并附有完整的答案。若你认真地操练这些问题，即在练习过程中仔细思考，写全答案，且在完成前不参阅正确答案，那么你就能自信十足地面对测验或考试，并能对一些更具典型性的答案的措辞作好事先准备。

我衷心希望本书有助于使研究方法和人类心理易于理解，并能使统计学不再神秘。我真诚希望该书给予那些着迷心理学却又经历常与“科学”相联系的可怕障碍的人以支持。我希望本书能帮你挽起袖子，投身于发现有关人们心理的实践之中。当然，我也希望该书还能让你在心理学课程中取得佳绩。

休·库利坎

谢 忱

我要感谢弗拉纳根(C. Flanagan),她在准备自己书稿时向我提出许多问题,帮助我厘清一些棘手的难题,从而在对这些问题的问答中发现了第三次修订版的撰写方式。也许这些夜晚的 E-mail 本身就能在某一天成为一本书! 我还要感谢伍尔夫(E. Woolf)、林肯(J. Lincoln)、米勒希普(S. Millership)、托马斯(S. Thomas)和汤普生(A. Thompson)在本书问世过程中给予的所有支持。最后,我要感谢在他们那里试用该教材内容的许多学生所给予的可能是不知不觉的帮助,使我得以精细其主题和完善许多撰写思想、策略和论证。

目 录

1

心理学与研究

本章内容

- ❑ 我们着眼于用**实证的证据**支持猜想的需要，理解科学之所为并非是“证明”理论的真实性。
- ❑ 我们着眼于心理学研究者用可替换的解释来挑战理论的方式，鼓励读者挑战“显而易见”的思想。
- ❑ 我们会看到公平、客观的研究方法是所有理论的基础，而且只有通过这种方法心理学家才能有一大批非同寻常且有趣的发现。正是这些发现使读者有可能喜欢阅读作为心理学内容的东西。
- ❑ 科学方法源于常识，读者会发现他们已经有足够的能力提出可能的方法来回答一个研究的问题。
- ❑ 我们开始着眼于科学的方法来询问本次研究所要评价或测量的是什么？研究的是谁？收集的数据有什么用处？
- ❑ 科学包括变量的观察和操作，它通常产生定量数据。我们还将评估近年来逐步增加的定性调查。
- ❑ 我们了解心理学家开展研究的各种原因：获得描述统计、建构和施测心理学量表、收集定量数据和验证假设。
- ❑ 我们着眼于出自常识但更富有逻辑的传统科学方法的历史、本质和发展阶段；提供一些示例说明相信常识的危险性。
- ❑ 为了推进科学发展，科学不应当总是用证据支持理论，而应当用证据挑战理论，尝试确立理论行不通的事实。
- ❑ 描述研究的整个过程，提供关于心理学研究是如何设计、审批、执行，直至（如果一切顺利的话）发表的简要介绍。
- ❑ 我们规定所有的心理学研究必须在英国心理学会制定的伦理原则下进行。

1.1　成为一个质疑者

你相信科学家研究中得出的所有结论吗？如果你回答“不”，那么就有了一个好的开端。当然，媒体经常报道某杂志得出了引人注目的结论（常常未经证实），但采访科学家时，科学家会显得极其谨慎。例如，前不久，报纸上刊登了一篇文章声称，心理学家可能已解决了“弗洛伊德之谜”（Guardian，10 September，2004）。文章充满了“巨大突破”之类的言词，声称研究者的工作“……足以证明理论的一部分，弗洛伊德提出的某些过程与人脑的特定活动有联系”。像我们后边将看到的，既没有心理学家，也没有其他素质良好的科学家说是“证明一个理论”。只有数学家证明理论，科学家用实验发现支持理论，并且试着指出有争议的理论在哪里不足以解释已知事实。实证性意味着是基于对现实事件的观察——简言之，事实。前面报纸报道的实际情况是，心理学家确认了大脑后部的一个区域似乎与梦有关，或者至少和梦的回忆有关。这个发现本身很重要，但肯定不能证明弗洛伊德的正确，也不能向我们表明梦实际上就是愿望实现的一种形式，只是向我们表明大脑的特定区域在梦的体验中很重要。仅此而已！



图 1.1　质疑者（‘ah, but’ thinker）

噢，亲爱的读者，那么本书会带出什么浪漫的东西吗？啊，某种程度上，是这样的——我们的意思是，如果我们判断所需相信的事物是否真实，仅是因为它们的迷人与美好，那就是出于“浪漫”了。我们这里更强调证据而非寻常所见，我们想知道的是，人们所声称的对人类的研究什么是对的，他们是如何得出结论的。这就是所谓的“科学”。事实上，你已经这样做了，因为你对奇怪的断言投去了怀疑的目光，也许会像这样说：“啊，他们是如何知道的？他们有什么证据？”

大体来说，当试图了解研究方法和心理学，你最好能养成某种令人气恼但又有帮助的质疑心态（“ah，but...” mentality，“啊，但是……”），就像前面提到的那样。例如，当你听说大量上网可能会导致抑郁，你会说：“啊，能否这样认为，抑郁的人更有可能将对互联网的使用放在首位？”换句话说，不断寻找**替代性解释**（alternative explanation），寻找研究设计中的缺点和限制，是会找到最终结论的。不断挑战！

无论什么科学，推动研究的是挑战和回答挑战的过程。见专栏 1.1。

专栏 1.1 亲爱的出租车司机，我们没有小费，但你脑筋不错!

一个由埃莉诺·麦奎尔带领的伦敦大学学院组成的研究团队引起了媒体的兴趣。他们声称伦敦的出租车司机大脑的特定区域(海马后部)比其他人有更厚的灰质(Maguire, Spiers and Good, 2003)。他们强烈表明，海马后部与导航技能有关，出租车司机由于围绕伦敦街区驾驶多年，使他们形成了更厚的灰质。支持出租驾驶导致灰质增厚的说法是出租车驾龄与灰质厚度存在强相关，驾龄越长，灰质越厚。然而就在这时，质疑声突然蹦出来问:

"但如果是该区域灰质天生就厚的人们发现街头导航更容易而从事出租驾驶，那又会怎样?"

这是典型的因果倒置。我们常常会颠倒因果方向。例如，不是打屁股导致儿童变得有攻击性，也许最初是由于儿童有攻击性，父母才打儿童屁股(我深信，打儿童根本不是一个好想法并且教会他们攻击——"你小，所以我能打你"。多好的示范!)。

在这种情况下，研究者期望这种反对，拥有更多灰质的人们更可能变成出租车司机。他们在非出租车司机组中检查了灰质数量与导航技能的关系。如果真的是灰质多导航技能就高，并因此成为出租车司机，那么我们会期望在非出租车司机样本中也存在灰质与导航技能的关系，但没有。灰质多与导航技能高并没有联系。因此，似乎是多年的驾龄导致了灰质增加。由于灰质多，导航能力就高而因此选择出租车行业的情况，但事实并非这样。

我们应当注意到这个研究赢得了"搞笑诺贝尔奖"(lg nobel prize)。该类奖项是因公布"首先是令你大笑，接着是令你思考"的发现而由科学幽默杂志《不可思议研究年刊》(*Annals of Improbable Research*)颁发的。在该杂志社的网站上(http://www.improb.com)你可以找到更多各种荒诞怪异的研究，包括在各种表面上拖绵羊所需力的精彩分析。

从专栏 1.1 中获得的重要信息是研究者用方法回答问题的方式。因为他们的设计结构良好，不是只坐在那里说:"我们认为出租驾驶提高了灰质是因为司机们的灰质比一般人更多。"他们设法表明，只有出租车司机才存在着导航经验与灰质数量之间的关系。

心理学历经一百多年的研究，获得一些令人惊奇叫绝的结论，就像梦所在位置的"发现"、海马与导航的研究。例如，我们知道:

- 许多人会不断地服从研究者下达的电击"受害者"的指令，使电击的强度达到明显会导致受害者死亡或严重受伤的程度。
- 在公开场合，许多人会从众于多数人的判断，即使多数人的判断是很荒唐的。
- 当自己犯了错误，人们倾向推诿外部环境("啊，是你没有告诉我!")，而把别人的错误归于他的人格或内在缺点("啊，对戈登的脑子你还能指望什么呢?")。
- 经历了某种脑手术后的人能够认识呈现在大脑一侧视觉系统的物体(如大象)，但要求他命名时，却给出呈现在大脑另一侧视觉系统的物体(如钥匙)的名字。
- 儿童不复制语言，他们尝试讲话看是否符合语言"规则"。

仅仅因为在这些发现的背后,研究者使用了普遍接受的方法,并且是以其他研究者能够理解和挑战的方式公布了结果,我们才有了心理学的知识体系……而且是那么的丰富!

现在让我们来弄清楚研究方法所用到的技术。当然,正如我在前言中提到的那样,许多人偏爱了解心理学家们所讲的内容,而不是了解他们通过怎样的方法才能够讲出来。然而,教学大纲编撰者明智地意识到,对任何一个心理学学生,学会如何做研究、如何认识设计的优缺点、如何避免不被骗人的结果和错误的统计方法所蒙蔽,是非常重要的。学习研究方法带给你的是技能,而不只是知识,并且这些技能在心理学课程之外也会令你受益。让自己沉浸在方法和统计之中,你就能向那些出于便利选择统计方法的政治家发出挑战,或看穿狡猾的广告所做的那些令人印象深刻的内容。学习研究方法帮助你加入心理学家的行列,能看到研究的优缺点而不是一味接受结果。

科学始于常识——我们都能做得到

如果愿意,每个人都能成为科学家。科学的逻辑和推理,不是只有受过良好教育的人们才能使用的技能(否则就太悲哀了)。我们都拥有这种技能。在你很小的时候,为了理解和预测周围复杂多变的世界,你就一直在使用这种技能。

为了使我们从真正的研究方法开始,让我们做一个练习来演示我的意思。在教授研究方法甚至统计学的时候,我始终坚信,读者对本书讨论的许多观点都已形成了基本的直觉概念。只是你可能没有意识到或详细思考,为了判断"大多数猫比波斯猫更喜欢选择 Moggy Munch(一种猫粮)",你必须说出十只猫中有多少只选择了 Moggy Munch。你可能会意识到调查样本偏差太大,思考样本如何才能更有代表性。换言之,你知道科学术语"公平检验"指什么。下面的方法向你展示你所做的。

假设你在学校食堂无意中听到下面的争论:

> "我不知道。所有战争你都听说过,而这些战争几乎都发生在炎热的国家。我猜炎热令人更有攻击性,更想突然袭击而不是事先预谋。"

虽然你已开始学习心理学课程而且假设想对人们的行为和思维提一些问题,但是我不希望你会接受这种粗浅的假设。我希望你应当做的第一件事是对这种断言寻找证据,而不是接受这种"安乐椅理论"(想当然理论——译注)。因此我要求你自己着手解决这个问题。你如何设计和进行一项研究来验证炎热令人更有攻击性的观点?也就是,为了收集证据支持或反驳这个观点,你会做什么?试着考虑几个你可能使用的替代性方法。

当要求执行这个任务,大多数人会突然地一怔,提出一些像专栏 1.2 的想法。我们都是潜在的科学家,我们都能想出某种试验或收集证据的方法。毫无疑问,你的某些想法与以下内容非常相似,或许一样。

专栏 1.2 验证炎热令人更有攻击性的观点的一些可能设计

1. 我们选一组人，让他们处于冷房间内讨论问题，并且观察他们是否变得具有攻击性。接着提高室内温度，然后再观察他们。
2. 我们选一组人在一个热房间内，播放暴力影片，要求他们说出电影中的人物攻击性有多大。然后选择另一组在冷房间内做同样的事。
3. 我们选择一组学生，在操场观察他们玩耍，看哪一对最有攻击性，哪一对次之，以此类推。我们把每对学生分开，形成攻击性相同的两组，然后我们让不知实验目的的观察者观看学生玩耍，一组学生一直在冷房间，而另一组学生在开始同样冷而后变热的房间，最后比较这两组学生的攻击水平。
4. 我们选一个小伙子，将他置于一个房间内，变换温度，每次要求他执行挫折任务，如用相对针眼较粗的线穿针。我们用问卷测量他的攻击性。
5. 我们在冰岛和墨西哥观察和采访交通高峰时段的司机。
6. 我们在一天里对某国不同气温地区取样，我们会要求那里的人们做一些事，如有人打了他们的小妹，他们认为该如何判刑或他们会做什么。
7. 我们跟踪一组男孩一年，记录他们在某周或某月里的所处环境的温度情况和打架次数。

1.2 心理学与研究

专栏 1.2 的每一个建议都是研究设计，也就是建立实验条件，收集数据，回答研究问题。研究问题，就像那个人在食堂所做的概括，“炎热令人更有攻击性吗?”所有的研究都会提出某种研究问题——研究者想发现的是什么。

心理学家如何回答研究问题

从根本上来说，他们做研究。他们决定做一个详尽的设计，就像专栏 1.2(尽管会更精确)，然后他们需要对他们的整个研究作出更多的决定：

1. 测量或观察什么? 如何准确地测量或观察?
2. 研究谁?
3. 我们如何处理收集到的数据?

决定 1 常常涉及对变量的准确测量。变量是可以变化的事物，需要在研究计划中准确定义。为了测量攻击性，我们必须精确说明选择什么作为攻击性测量的工具。例如，我们可以使用问卷，也可以对每小时内的已定义的攻击动作进行计分。变量产生定量数据，也就是数字形式的数据。我们将在第 2 章论及变量的测量。注意到在专栏 1.2 的 4、7 给出了测量“攻击”的具体方法。

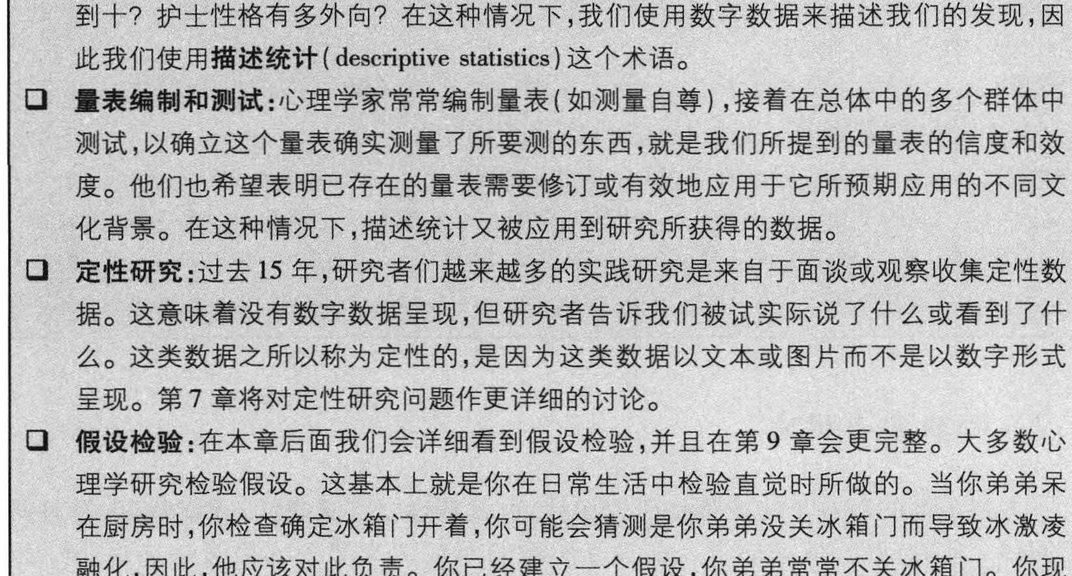
另外，决定 1 也包括某种记录定性数据的方法——非数字化信息。这可能由面谈中的所有对话构成。定性数据及其记录方法将在第 7 章论及。

决定 2 涉及研究中的被试。例如，像专栏 1.2 的例 1，在每个条件下测试相同被试有什么优缺点? 相反，例 2 使用不同组类的被试有什么优点? 问题又是什么? 大量的心理学研究使用学生做被试，而且大部分学生主修心理学专业(Banyard and

Hunt,2000)。这意味着,把学生结果延伸到其他人群,公平吗?这些问题涉及"研究谁、哪类群体以及如何确保取样公平",将在第2章论及。

决定3也许最难,令你在学习做心理学研究中承担最大的压力。为了解决这个难点,让我们看一些不同类型的研究,它们能产生不同种类的数据,因而需要使用不同的统计分析。见专栏1.3所给示例。

专栏1.3　心理学家在他们的研究中做什么



- **描述性研究:**有时我们只想弄清情况,婴儿何时能准确识别母亲?多少三岁儿童能数到十?护士性格有多外向?在这种情况下,我们使用数字数据来描述我们的发现,因此我们使用**描述统计**(descriptive statistics)这个术语。
- **量表编制和测试:**心理学家常常编制量表(如测量自尊),接着在总体中的多个群体中测试,以确立这个量表确实测量了所要测的东西,就是我们所提到的量表的信度和效度。他们也希望表明已存在的量表需要修订或有效地应用于它所预期应用的不同文化背景。在这种情况下,描述统计又被应用到研究所获得的数据。
- **定性研究:**过去15年,研究者们越来越多的实践研究是来自于面谈或观察收集定性数据。这意味着没有数字数据呈现,但研究者告诉我们被试实际说了什么或看到了什么。这类数据之所以称为定性的,是因为这类数据以文本或图片而不是以数字形式呈现。第7章将对定性研究问题作更详细的讨论。
- **假设检验:**在本章后面我们会详细看到假设检验,并且在第9章会更完整。大多数心理学研究检验假设。这基本上就是你在日常生活中检验直觉时所做的。当你弟弟呆在厨房时,你检查确定冰箱门开着,你可能会猜测是你弟弟没关冰箱门而导致冰激凌融化,因此,他应该对此负责。你已经建立一个假设,你弟弟常常不关冰箱门。你现在有证据支持你的假设。但你知道你的假设可能不正确。因为有可能你弟弟仅此一次,而实际上你妹妹是罪魁祸首。在心理学研究中包含相同的逻辑程序,但收集的数据量当然会更大,分析也更科学。证据常常以统计形式收集,例如通过计算在冷热不同温度条件下正确回忆词汇表中的单词数,来帮助我们确定两者差异是否显著。**统计显著性**将在第9章详细解释,但现在我们只是说我们需要弄清楚,是否数据表明温度真正影响了记忆,或者是否不同条件的差异过小推论不够严谨。在上面的例子中,你把你弟弟一次事件概括到了你一般的假设。在心理学研究中,我们称为**推断统计**(inferential statistics),以评估我们发现的统计概率只是一次侥幸还是更一般效果的证据。从现在起,当我们谈论"效果",是指变量间差异(如热房间的人比冷房间的人更热衷于争论)或变量间关系(如越热,你喝水越多)。

练　习

1. 云能成为变量吗?
2. 什么是研究设计?
3. 定性数据与定量数据有何区别?

答 案

1. 能。任何物理或心理特征都可以。例如,当天空中有许多、一点、几乎没有云时,我们能测量人们的心境。但我们必须仔细定义"一点"和"许多"的差异。
2. 研究执行的方法就是研究设计。
3. 定性数据是非数字的,包含有意义的信息,如文本或图片;定量数据是对测量的记录。

关键术语

描述统计(descriptive statistics)	定性数据(qualitative data)
实证(empirical)	定量数据(quantitative data)
假设(hypothesis)	研究设计(research design)
推断统计(inferential statistics)	研究问题(research question)
被试(participant)	变量(variables)

1.3 为什么都必须是科学的

有一个普遍错误的观念:科学就是发现事实。实际上,一旦科学发现了事实,这些事实对科学家就不太有用了。它们对工程技术人员、医生、建筑师和养育员等一定会有更多用处,因为他们在日常工作中使用和应用这些知识。科学是关于发现和解释。一般来说,我们知道,儿童从 9 个月起,如果他们的照看者突然不在了,他们就会变得不安;从 18 个月起可以表达粗糙的语法;到 6 岁为止他们一直相信无生命的物体也有某种生命,诸如此类。但是心理学家想知道的是,他们如何这样做?更重要的是,为什么?像其他科学家一样,心理学家发展理论并且验证理论能否作出解释。

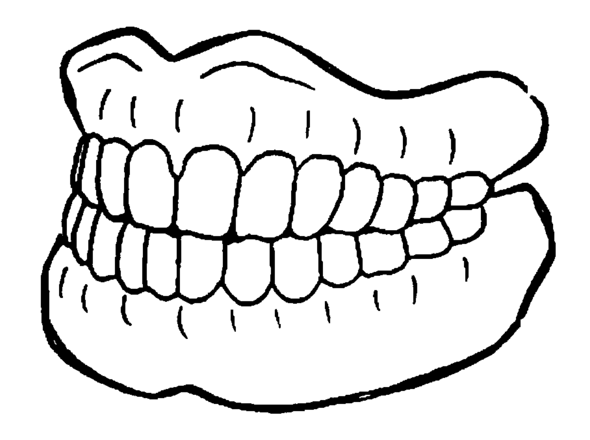


图 1.2 数一数有多少颗牙齿

在 17 世纪弗朗西斯·培根(Francis Bacon)之后才出现这种自然科学,并随之出现经验主义者,他们引进了全新的建立知识的方法。在这之前,关于世界的知识受到传统思维的束缚,这种思想可追溯到 2000 年前亚里士多德时代。为了让你品味那个时代的思维有多么不同,来看看罗素(Russell,1976)所列举的例子。亚里士

多德牙齿理论的逻辑结论是:女人牙齿比男人少。但他没有努力去检查,即使他有几个妻子可以作为研究对象。在亚里士多德时代,如果"事实"来自逻辑推理或者符合世界及其缔造者的主要宗教观念,它们就是科学的。也就是说,亚里士多德的世界观导致了他的逻辑结论——女人牙齿应该比男人少。根本没有必要进行**实证**检验。这就是我们为什么需要科学:阻止人们作出各种荒唐的声称,阻止其他人只是因为"对自己有意义"或"与自己经验吻合"而相信这些声称。

1.4 科学方法

因为在17世纪,哲学家对亚里士多德思考世界的方式变得不再满意,他们很快转向科学调查观——尽管在今天大多数人看来是理所当然之事。即使在学校里憎恨科学的人,也肯定会立刻反对亚里士多德荒唐的、毫无根据的论题。我们会说:"啊,你数过吗?"

表1.1展现了普遍接受(但相当近代的——只有几百年)的关于当今科学家如何思考和调查的模式及其一些实例。

表1.1 科学方法——或多或少

步骤	方法举例
1. 进行观察(使用测量分类等)	注意个体在攻击数量上的变化
2. 探查和总结数据的一般模式	观察受到体罚的儿童攻击性更强
3. 为模式(理论或假设)提供解释	儿童仿效具有攻击性的父母的行为模式
4. 通过实验或进一步观察检验预测,进行假设检验	在瑞典,观察1979—1999年,禁令惩罚的结果。卷入刑事犯罪、强奸、自杀及其他法律效力的青年显著下降(Durrant,2000)
5. 支持或挑战理论	支持惩罚—导致—攻击理论
6. 对上面的挑战进一步寻找支持或继续挑战	也许在其他国家同期有相同的影响(也许没有)

这种方法荣幸地获得了一个复杂的名字:**假设—演绎方法**(the hypothetico-deductive method)。我们提出理论和假设,然后检验理论,演绎支持或挑战理论。

当然,大部分研究并不是顺利地遵循这些步骤。从进行观察到检验假设,绝不会是一条直线。常常是,我们已经知道了许多有关"事实",但我们想调查理论"为什么某种行为会发生?"我们不会只等着观察数量的增加。因此,方法中的前两步常常是理所当然的,从第三步开始,我们努力解决什么应对人类的某种行为负责。我们会直觉地认为某个变量可能会影响行为,并且通过研究来调查。

在需要解释的整个知识范围内,一个开始收集数据的经典例子是,调查从1960年后期以来的助人行为。起因是1964年发生在纽约的凯蒂·热那亚谋杀案,令人毛骨悚然。38人知道有一妇女在他们的公寓外面临危难,妇女的哭喊声令人难以遗漏这样的事实,妇女多次被刺,后被强奸。但是,只有一个人叫了警察,那已是攻击发生后的半小时,当时妇女已经死了。达利和拉塔尼(Darley and Latané,1968)并没有等待收集许多关于帮助行为的观察。他们开展了一系列实验,研究几个可能对这种非人性的消极行为产生影响的变量,诸如他人在场的数量、现场情况的不明朗、

他人行动的缺少、个人安全,以及其他因素。就这样,他们从这样一件特殊而又令人痛心的事件出发,发起了一项新的但现已有多年历史的助人行为调查。

1.5 科学方法的规则

上面提纲列出的科学方法步骤,在研究者实行研究时,并不总是处于研究者头脑最前沿的内容。相反,如果研究是科学地进行,那么至少必须根据下面的原则执行。

- ❑ 仔细定义测量,并尽可能准确地操作。
- ❑ 所有数据的收集尽可能客观。
- ❑ 所有测量、方法、设备和收集的数据通常都以学术期刊上发表论文的方式获得公开有效的检验。
- ❑ 进一步的细节,按要求可从研究者处获得。
- ❑ 无论多么令人不舒服,都可以接受所发现的事实(但未必都是它们暗含的东西)。
- ❑ 研究者的人格与他们所发现的东西是否被认真对待无关。重要的是他们的工作内容。但是,根据细微证据作出宏大的断言或者篡改结果是要受到批评的。
- ❑ 常识性假设总是受到挑战,必须拥有实验证据支持。

尽管我在前面暗示过,在你的日常生活中应用了某些科学原则,并且为你提供了开展科学的概念,但并不等于说日常思维总会这样。常识的麻烦在于,它走了捷径但假设太多。比如以一个"显而易见"的假设为例,由于感冒在寒冷和潮湿的冬天很普遍,因而寒冷和潮湿是感冒的直接原因。但事实并非如此,感冒是由病毒传播的,冬天我们常常在更拥挤的室内,因而病毒更容易在人群中传播。寒冷和潮湿只是感冒的间接原因——可以穿得暖和些,但温暖不能阻止病毒。

还有,常识思维常常并不能足够谨慎地探究到底。考虑下面的情形:

1. 如果我再把一张纸折叠几次,比如 4 次,你会看到纸变厚了。如果我把纸折叠 50 次会有多厚?现在请认真考虑,大体估计厚度,比如厚度可达到桌子、墙、房子……
2. 男女相比,哪个性别对其同伴更有身体上的攻击性?

问题 1 的答案是高度可直达太阳,并加上返程的一半!大约 1.35 亿英里!

对于问题 2,根据亚瑟(Archer,2000)的研究,女性更可能具有身体上的攻击性。但在你向我咆哮我是个大男子主义者(sexist pig)之前,注意到男性更可能导致身体上的伤害或使他们的女性伴侣受伤。在亚瑟的研究中,女性只是更频繁使用身体上的攻击(通常是非常温和的)。而且研究是在美国进行的,之所以引人注意是由于结果如此令人惊奇。但这个研究告诉我们,在我们认为假设正确之前必须检查,即使是显而易见的假设。

心理学家设计了许多测验表明,大多时候人们并不是富有逻辑性的。他们并不按大多数人所声称的在决策时使用严格的逻辑行事。下意识的、情绪性的或主观条件反应常常起支配作用,这种情况对男性和女性同样适用。里托威和巴罗(Ritov and

Baron,1990)问被试一个假设问题:想象一下,假如有一种流行感冒,儿童死亡率为万分之十。有一种疫苗肯定能阻止疾病但也可能会导致死亡。他们要求被试决定自己的孩子接受疫苗后的死亡风险的最大程度。被接种后死亡的风险平均值为万分之五。换言之,被试选择了拒绝疫苗而侥幸逃过流行感冒的风险是接种后风险平均值的两倍。这里发生着某种"不可思议的思维"。不知怎的,给予儿童疫苗的积极行动略微有死亡风险,比让儿童被动死去更可怕、更犯罪,尽管后者灾难概率会加倍。与英国近年来发生的情况很相似。当时关于联合疫苗注射可能是诱发孤独症原因(Shimizu and Rutter,2005)的微弱证据(实际上是难以相信的),令父母避免让儿童注射疫苗,从而可能导致麻疹、腮腺炎和风疹病例数量的激增(Jansen et al., 2003)。

为什么我们不能说我们证明理论是对的

我早些时候说过,科学家常常不谈论**证明**理论。我会尽力解释其中的理由。现在让我们看看假设检验,以及如何建立经得起时间考验的理论。试一下这个练习:

> 我正在思考一个可以产生有效的数字集合的规则。例如,如果我的规则是递减奇数,那么9、7、3或7、3、1是符合的,但9、8、5则不行。接着,想象我头脑里有这样一个为下面数字适合的规则:2、4、6和8、10、12。
>
> 假设我现在要求你产生更多的数字集合,我会告诉你数字是否符合规则。你的任务是一直这样做,直到你能确定这个规则为止。下一个数字集你会试哪一个?

你很可能会选择等差递增数列,如14、16、18或者3、5、7。你想过一组数字作为可能的答案以符合你头脑中的规则吗?我这样问,是因为除非你找到一组**不符合**的数字,否则你哪儿都去不了。如果你认为数字递增2,那么问3、5、7是否符合有什么意思?如果符合,我会说"是的",但你不知道规则是"递增2"还是"偶数"或者只是"递增"。如果你想检验一个预测,那么试一组不符合你预感的数字。因此为什么不试1、3、7呢?如果对了,你可以肯定地排除"递增2"和"偶数"。接着,你转向另一组数字排除另一个可能。如果我们认为新规则是"递增",那么我们尝试3、2、1之类。我的规则实际上是"递增"。如果你不想保留秘密,在你朋友身上试一下。

在科学上使用这种"挑战规则"系统,我们就能排除许多可能的理论,其余剩下的理论更可得到稳固的确立。非常多的科学和心理学研究被设计来挑战假设——一个可能的解释,以期排除它。不能证明它正确,我们通过排除可能的解释来支持它。

科学家不是证明理论而是挑战理论吗

利用表1.1的提纲,研究者提供证据支持理论而不是证明理论。智力通过基因由父母向下传递,这个理论有什么证据?同卵双生子拥有相同的基因,因此研究者检查配对双生子的智力(IQ)。确实非常接近。这就证明了智力是天生的吗?啊,不,绝不。同卵双生子拥有相同的性别、年龄,生活在相同的环境,并且一般在相同的家庭、经济、文化环境被相同的父母抚养。同卵双生子的IQ接近,既是先天观的证据,也是环境发展观(后天观)的证据。当面对这种模糊的证据时,科学家寻找一个研究设计来排除其中一个观点,就像我们在上面数字排除时所做的一样。会是什

么呢？啊，如果智力主要是抚养的结果，那么在不同环境成长的同卵双生子不该有相似的智力水平。相反，到目前为止，已经获得的研究表明确实是这样。但这绝不是终点。我们不能用同卵双生子进行实验——见第2章。我们只能研究那些不幸被分开的孪生子，抚养在相当不同的家庭里，并且同意接受测试。令我们感到困难的是，寻找在教育机会、家庭态度、经济状况等不同环境被养大的孪生子。如果被分开，大部分双生子被置于在这些方面相当相似的家庭里。现在我们不必对这个例子作进一步诠释，但你可能会注意到，就这个问题心理学历史上曾有过最激烈和最残酷的争论。

现在你可能想起其他研究设计，能为先天-后天问题提供证据，见专栏1.5中的一些建议。

发现和结论

这里强调的是，没有证据“证明”理论或者理论是清晰确定的。总有质疑问题，总有另一种方法解释发现。研究者在研究中呈现的是他们的发现——他们收集、总结、分析的数据。他们接着讨论这些发现，并提出一个结论。这个结论解释了证据如何支持（或不支持）研究者最初开始调查的理论——见专栏1.4。就像你在智力的有关例子中看到的一样，许多研究背后的动机是，试图**排除**以前发现的可能解释。毕竟，这只是正常的日常思维的拓展。

专栏1.4 发现和结论

总是要仔细区分**发现**和**结论**。**发现**是在研究中真实发生的东西——也是结果。结论是研究者在背景理论下考虑结果时所作的推断。例如，同卵双生子的IQ相关性相当高，这是一个发现。从这个发现，研究者可能得出结论，遗传在智力发展中影响很大。相反，就像上面看到的，这并不是唯一的结论。也许，与一般兄弟姐妹或非同卵双生子比较，双生子共享了如此多的共同因素，因此他们显得如此相似。我们之前早看到，出租车司机比非出租车司机趋向于有更大的海马。这是一个发现。它呈现了研究者发现的事实。**结论**是，拥有大量关于伦敦道路系统的经验导致脑部特定区域增加。发现应当总是清晰、不含糊的，如果可能，也几乎没有争议；相反，结论常常充满争议。

专栏1.5 调查先天-后天问题的方法

怎样比较收养儿童和养父母呢？如果基因主要影响IQ，这些收养儿童智力一般不应当与养父母相似；但如果养育是关键因素，他们的智力就会相似。我们可以看看收养儿童与养父母的关系，然后再把儿童与亲生父母的关系进行比较。结果是儿童与亲生父母更相似。支持了先天观吗？是的，但还有另一个质疑。一般来说，养父母来自其他的社会阶层，这种比较并不公平。我们所需的是利用相同的父母进行比较。让我们比较养父母与他们的养子，然后把同一父母与他们的亲生子女比较，两个关系的差异不显著。我们也可以试着比较，收养儿童和他们的养父母的关系，以及相同儿童与他们的亲生父母的关系。每一个设计都有自身的问题，产生的证据会受到挑战。

寻找挑战性检验,而不是确认性检验

考虑以下的思维训练:

"我肯定,是人造黄油把亨利皮肤弄得满是疙瘩。""我不这样认为。怎么可能呢?那好,我们别给亨利吃人造黄油一段时间。如果他还起疙瘩,那就不是人造黄油的缘故。"

上面的检验符合我们提纲里列出的原则:持续给亨利吃人造黄油,无法产生更有力的证据。试着停止麦淇淋,看疙瘩是否持续。如果回过头看前面的有关出租车司机研究,你会看到,研究者也需要设计某种检验来排除可能性——出租司机有更大的海马,是因为他们生来海马就大才导致他们从事出租行业。导航技能越高,海马越大,这种关系在非出租司机中不存在。但这种关系在出租车司机中存在。因此,我们似乎可以排除假设——生来脑子越大就越驱使你开出租车,并且进一步支持假设——出租驾驶发展了大脑。

练　习

1. 科学方法在什么情况下与"常识"相同,在什么情况下与常识不同?
2. 科学方法有哪些主要阶段?
3. 什么是发现?什么是结论?

答　案

1. 两者包含的逻辑思维是相同的,但科学方法会使我们更彻底地检验,不轻易接受表面上"显而易见"的理论。我们考虑各种可能的替代性观点,并且通过进一步的实证结果尽力排除这些观点。
2. 见表1.1。
3. 发现是研究的清晰的结果——是对显著性的差异或关系的总结性数据分析。结论是研究者认为发现的结果支持或挑战某种理论的手段。

关键术语

结论(conclusions)
科学方法(scientific method)
发现(findings)
理论(theory)
假设-演绎方法(hypothetico-deductive method)

1.6 心理学研究到底做什么

开展实验或其他研究常常只是整个研究过程的一小部分。实际上,我们可以把研究看做没有终点的循环(见图1.3)。

心理学家为什么进行研究的部分原因已列在专栏1.3中。通常从已建立的理论或从发生在我们周围新的或重要的事件出发,通过近期研究分析形成研究计划。为了验证,研究者可能会重复他人已做过的研究。典型情况是,他们会重复一个研究以排除可能的解释,例如同卵双生子抚养或只研究养父母(他们自己有孩子)。

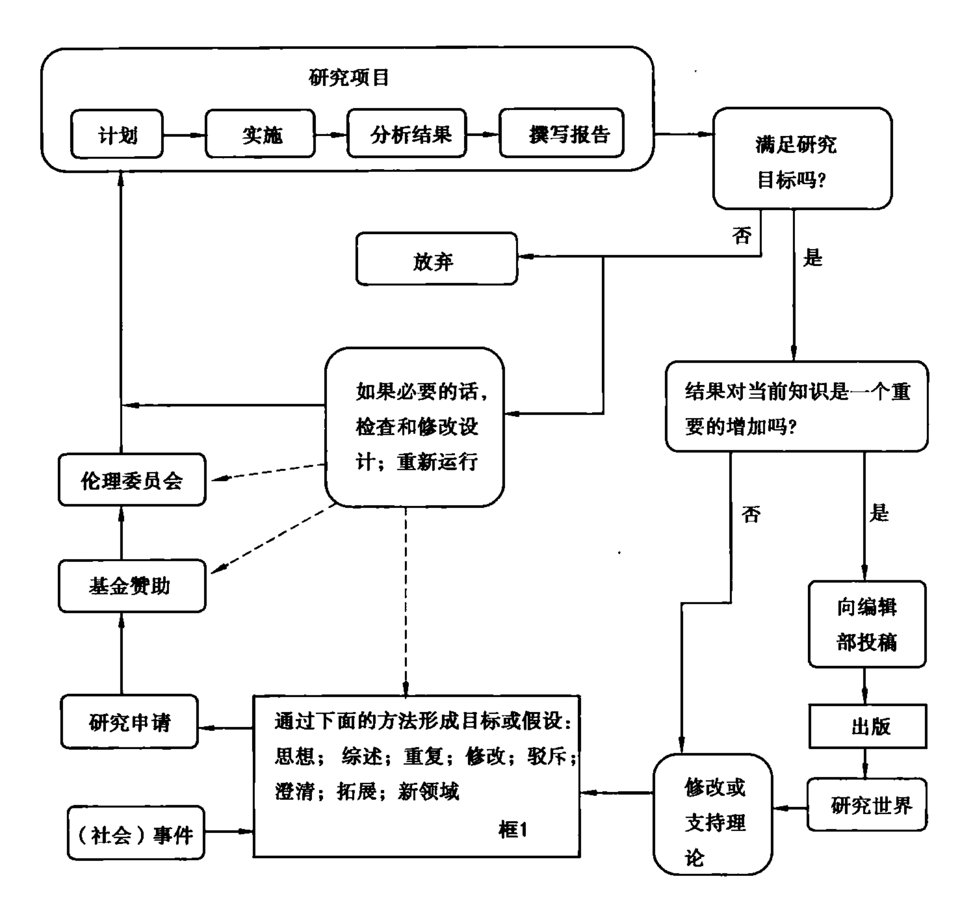


图 1.3 研究的整个过程——从框 1 开始

研究者也会把一个研究结果拓展到更多的群体或其他情境。例如,许多研究表明,咖啡因常常以药片形式作为实验控制药物,它能抵消疲劳的影响。但一般的高能饮料也能起同样的效果吗?瑞纳和霍恩(Reter and Horne,2002)研究表明,一种含有咖啡因的商业饮料——红牛,对健康的青年人有同样的效果,这些人前一晚被剥夺了睡眠,从事一项单调的模拟驾驶任务。这个结果向我们表明,一般饮料也能降低疲劳和驾驶失误,但我们还可以把这个研究进一步拓展。咖啡因饮料能降低现实生活中汽车驾驶的疲劳吗?多少量的红牛饮料可抵消睡眠剥夺的影响?某种程度上,研究就是**发现**某种影响的局限性。

真正的新研究

每隔一段时间,研究者就会研究完全崭新的现象或开辟新领域。甚至关于出租司机和梦的研究,曾被出版物较早提及而引发轰动,被称为“第一”,实际上这些研究是作为一般研究项目——大脑是如何组织特殊技能和经验而开展的。相反,1964年发生在纽约的凯蒂残杀案,引发前面曾提到的达利和拉塔尼进行旁观者干预研究。搜索大量刊载心理学文章的电子数据库,发现从 2000 年 11 月 9 日到 2004 年末,提到“恐怖主义”或“恐怖分子”的研究文章,是 1872 年有记录以来包含相同关键词的所有文章的两倍。

整个研究过程

研究过程不只是让被试做些实验。在图1.3,从第一个方框起,心理学家从我们刚刚讨论过的任何来源,都能形成新的研究方案。下一步是寻求资金来资助项目。大学教师工作的一部分就是开展研究。但是,除了时间之外,许多项目需要设备、问卷、研究助手等来测试被试。资金来源包括:大学或医院研究经费、中央或地方政府、欧盟研究组织、私营企业、慈善机构、私人捐赠者(偶尔)。这些组织、个人或研究者的老板必须对研究价值满意,即对他们、社会或一般的科学知识积累有益。通过准备研究方案寻求资金,是研究者工作的很大一部分,有时相当枯燥。

当研究者的目标与提供经费的组织,其活动内容不一致的时候(例如,大型的烟草公司),赞助就变得很有争议了。这是心理学家们必须作出决策的难点,注意自己的道德原则,但更重要的是注意约束他们的伦理实践原则。在英国,这些原则由英国心理学会(BJS)制定,同时该协会也提供在研究中如何根据伦理对待被试的一套指南(BPS,2000)。大学研究工作需得到伦理委员会同意,他们也考虑其他事宜:被试体验压力或不舒服的程度,欺骗或信息保留是否必要,导致被试羞耻或尴尬的程度,等等。在第 13 章,我们讨论心理学研究中的伦理问题和原则。

在研究设计完全开始之前,研究者可能需要预实验,以确定材料和程序是否可行,以解决任何未曾预见的困难。

在预实验里,实验或测试只在几个被试身上进行,但数据不能用在最后结果中,可以用来调整测量。例如,在关于咖啡因和记忆的研究中,从**预研究**中可能发现,要显示任何咖啡因将导致记忆力提高是十分困难的,因此记忆任务需要变得容易一点(或者咖啡因的剂量可能需要再增加一些)。

图 1.3 中方框“研究项目”包括了本书的大部分内容。在你的心理学课程“研究方法”也会经历这些阶段。就像你能看到一样,对专业心理学家来说,这只是整个过程的一小部分。研究计划还包括:准确设计、材料、程序、获得设备、寻找被试、实际数据收集的统计。

如果研究实现了目标并且对某个领域的总体知识有些贡献,研究者就可以向专业研究杂志(如《英国心理学杂志》(*the British Journal of Psychology*))投稿。杂志编委会就会邀请审稿人(学术同行)进行评审。这些审稿人并不知道作者的身份。如果评价积极并且后来被接受,那么报告就可以发表了。

无论研究报告是否被出版,如果仅仅是作为有前景的研究被证明无果或预期结果被证明失败了,它们对理论的发展都可能会有一些小的影响。如果发现挑战了已存在的理论或研究方向,那么理论的创始者就会对研究的设计和程序提出质疑,以维护自己的理论而反对明显的挑战。他们可能会向调查者索要原始数据——原始未经加工的结果(如完整的问卷)。一些研究者会重复研究,另一些则修改研究……我们又回到了整个研究过程的起点。

练 习

1. 合上书不要看,尽可能地列出所有的研究阶段,从想法萌发到出版以及对当前理论的影响。
2. 给出进行预实验的理由。从预实验中,研究者可能发现什么?

答 案

1. 见图 1.3。
2. 见 1.6 节讲到的内容。

关键术语

伦理(ethics)	预实验(pilot trials)
杂志(journal)	原始数据(raw data)
预研究(pilot study)	

2

测量人类

——变量和样本

本章内容

- ❑ 尝试测量人们的行为和心理特征。
- ❑ 探寻什么是变量及什么是心理学构想?
- ❑ 探寻变量是如何严格定义的,以便其他研究者可以对相同的概念进行操作并对相同的假设进行验证。
- ❑ 心理学研究中接受测试的群体被看做总体的一个样本,因此我们的样本必须能够代表总体,才能把结果推广到总体中去。
- ❑ 介绍了几种抽样方法。有的基于随机选择和最小化抽样偏差,有的本质上是随意的,还有的是带有目的性的——收集小群体。

在讨论心理学家是如何调查有关人们的看法之前，我们必须考虑两件重要的事情：

1. 心理学家如何构建日常（和非日常）心理学概念的测量，如智力、依赖或记忆。
2. 心理学家如何确定测量对象。

探寻这个棘手的测量领域，需要求助于看似愚蠢但很具体的日常例子。这些例子不全是心理学的，但我希望学习者们能明白心理学家调查人类时所使用的方法并不比调查一般物理世界的那些大家所熟知的方法更神秘。

设想你搬进了新居，想勘察一下后花园草坪的健康情况。你可能会从一个邻居那儿获得微妙的暗示。她一边点头跟你打招呼，一边嘟哝"哎，多好的房子，可草坪就像垃圾场！"

首先你需要决定测量草坪健康情况的标准是什么。假设选择草坪的生长速度作为测量指标。我们把生长称为变量，需要对其测量指标作出明确的规定。你可以选用两周前后叶片长度的毫米数为测量指标（考虑到不想使两次修剪的时间间隔过长，从而让新邻居认为你是一个邋遢的人）。当然，草坪的健康状况还可以选用其他更多的指标来衡量，但我们为了作比较就姑且把生长速度作为健康状况的指标吧。我们称之为草叶生长变量的可操作性测量——正如本章第一部分看到的，探寻心理学家如何测量变量。

接着你需要做的是决定测量哪片草坪的叶子。显然你不会给整个草坪都做测量！评估草坪的生长至少有两种方法可用。你可以选择一个叶片为样本，在两个星期的这个时间段的前后分别做测量。另外，你也可以在两星期这一周期开始时选择一个样本，结束时选择另一个等价样本做测量。当然，这里的问题是如何选择草叶的代表性样本。不该选择邻居家猫常撒尿处叶子发黄的那片草坪，也不该选择七叶树下稀疏的小叶做样本。本章第二部分的主题即如何从总体中选择等价的和有代表性的样本。

2.1 变量——心理学里被测量的事物

变量和测量

研究的本质就是观察变化，对象可能是鸟儿、地质，也可能是情绪。如果没有变化，就没有观测的对象；如果所有鸟儿用相同的标志，那可能是同一物种，如果鸟儿不动，我们就不能描绘他们的迁徙、择偶和食物偏好。

如果事物发生变化，它就是可变的，科学上我们称之为**变量**（variabe）。下面是一些熟悉的变量：

身高	——随年龄增长而变化。
时间	——例如，完成拼图游戏。
政治党派	——不是可度量的单位，而是类别。
情感	——对父母、同伴或兄妹。
态度	——对恶霸、猎狐或酒精滥用。
焦虑	——紧张的行为、内心恐惧的感觉等。

上述方框中的一些变量是很容易测量的。有趣的是，正是心理学的变量——情感、态度、焦虑才是最难测的。

> **攻击**是心理学家研究的非常多的一个变量。我们再看看第1章专栏1.2的内容。所给的例子中哪个对“攻击”做的定义最清晰呢？

专栏1.2中例1、例3、例5未对攻击作出定义，因此我们不知道在这些研究中攻击是如何被测量的。例2稍微好一点，但我们既不知道“提问”的方法，也不能准确地知道被试是如何回答的——是开放性描述还是通过一份问卷？例4指出是使用了一份问卷，但仅此而已。例6建议从一个容易测量的开始——我们要求被试对一个虚构的罪犯判刑。研究假设他们感受到的攻击性越强，他们所判刑的刑期就越长。但第二个建议有点令人迷惑。我们如何量化（例如，给一个数值）他们的反应，如“去敲碎他们的头”或“砸他们的车”？是否其中的一个比另一个更有攻击性？尽管例7中我们首先必须确定什么是真正的“打架”，它所提供的测量还是最准确的。

心理学构想

攻击性存在多种测量方法的事实让许多心理学初学者认为，假设有如此多的方法都可以来测量攻击，那么攻击到底是什么则是不明确的。这种观点假设总有一个“正确”的方法能够测量攻击。然而答案是，没有！更糟的是，这种假设认为在世界上的某个地方一定有一种东西是“真正”的攻击。这种现象值得我们深思。

温度的概念是个很好的例子。你在任何地方都不可能找到温度，也不能找到压力，但温度和压力是确实存在并能被准确测量的变量。原因在于温度和压力都是一**种状态**。你只能说温度高了，因为发生了某些事情，如你觉得热、冰溶化了、洗过的衣服比寒冷时干得更快。你只能说出攻击的状态，因为发生了某些事情，如汽车被砸了、人被撞了、有人说脏话。这些类比可能不够完美，但这种思维方式对我们是很有帮助的。

众所周知，攻击是一个心理学构想。我们是如何获得这些构想的呢？孩提时，我们就是小心理学家了：我们观察行为，当他们形成系统时我们便学着给它命名。因此，我们知道了妈妈什么时候生气（如果我们不知道，上帝也会帮助我们），后来，我们又了解了焦虑有什么征兆。当然，这就提出一个疑问，即在我们内在的心理自我中是否存在某种被定义为焦虑的东西。对此，生理心理学家强调自主神经系统活动，而认知心理学家则更聚焦于恐惧思维。

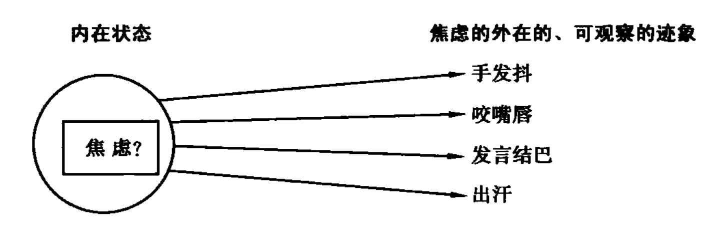


图2.1　内部假设构想的外部标志

在此，我们没有足够空间在哲学层面上充分讨论焦虑和攻击到底是什么，它们

是否真的存在。这里需要强调的是,由于我们意识到理解攻击和焦虑的方式如此之多,那么采用几种不同方法来测量它就不足为奇了。

心理学构想的黑箱法

我们所测量的东西是否真实存在并构成一个问题。从某种程度上说,心理学家与其他科学家做法相同——使用“黑箱法”。他们声称:“我们不能直接观察 x(如夸克),但如果它存在,那么 X 就会发生。我们检验事实是否如此。”如果事实果真如此,那么关于夸克的理论就得到了支持。在心理学领域中,我们把人类的心智视作一个黑箱。我们或许永远也看不到它的内部,然而我们能够假设心理学构想是以某种方式连接在一起的,并能用我们所能测量的东西来验证它。图 2.2 表明了这样一种可能的排列。

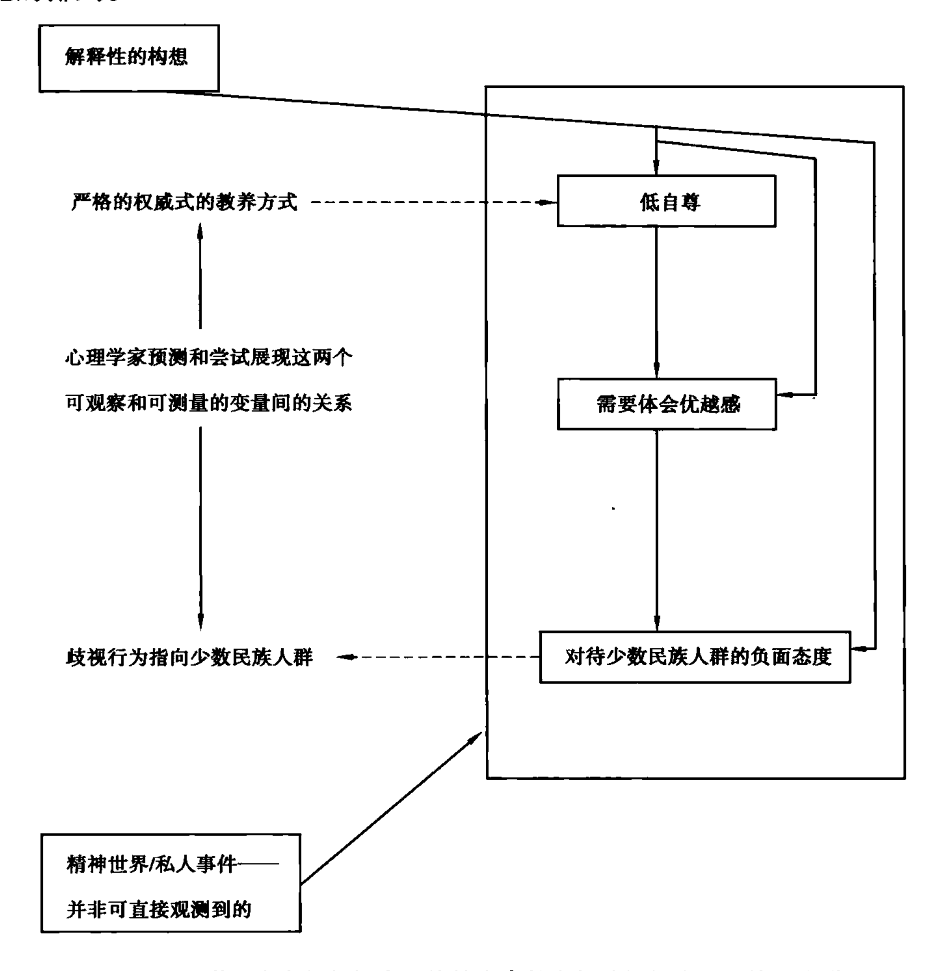


图 2.2 心理学研究者如何提出严格的家庭教育与歧视行为之间的理论联系

一个研究者可能假设,对少数民族的歧视行为(我们可观察到的东西)可能是由对少数民族强烈的内在态度(我们不能直接观察的东西)所引起的。反过来说,这种态度是由低水平的自尊感引发,从而产生了体验优越感需要。是什么导致了低自尊感呢?成长过程中受到欺负和情感匮乏,也即专制型教养方式是导致低自尊的原因。家庭教养方式则是我们所能够观察的,至少我们能够客观地测量父母所使用

的典型的教养风格。通过观察是否父母的教养方式越专制儿童越倾向于产生歧视行为，就能验证我们的理论了。

操作性定义

这里作个总结。我们可能会对内部状态和构想产生疑问，例如人格变量的外向性。但我们能够测量通过经验获得同特定构念相联系的这些状态的外部表现。也就是说，我们可能不清楚什么才是真正的“外向性”，但我们确实可以测量到它的外部表现，如对事物比想法更感兴趣，渴望遇见新的人群。

现在的问题是我们如何准确测量这些外部表现。专栏1.2中的每一项研究都需要对攻击有一个严格的定义。需要记住的是其他研究者可能会重复我们的研究，他们需要准确地知道我们是如何测量攻击的。心理学家们喜欢称之为操作性定义。变量的操作性定义是指测量一个构想所需要的一系列行为，如攻击。从某种意义来说，特定的研究构想是由其测量方法来定义的。例如，在专栏1.2的例7中，给出了一个精确的定义——一个月中的打架次数。这显然并不是攻击的全部内容，而只是在研究中采用的一个强有力的测量标准。

定义：挑剔但正确

希望每次在给学生的作业上写“准确定义你的变量”或类似的批语时，我都可以获得一英镑，甚至是一欧元。学生最初也认为给变量下操作性定义很难。下面是一些摘自学生的报告中的句子，可以看出他们对变量的描述都是很粗糙的。这里的“粗糙”是什么意思呢？让我们想象一下，你想要重复他们的研究时，该如何测量这些变量呢？基本上，你必须向研究者做不必要的询问。

“……在炎热房间条件下攻击性更高。”
“……训练组将有更好的记忆。”
“……男性在态度上更大男子主义。”
“……被试将认为他们比父母更聪明。”

导致操作定义“粗糙”的原因可能是受日常生活中常常使用粗糙定义的影响。例如，设想你的导师在早期的心理课程中展示过如何轻松地利用名字韵脚的方法来匹配名字和面孔（如想象一个扳手拧开了一个人的牙齿，这个人就叫汉纳（Hannah，与扳手 Spanner 谐音），现在她要求你对结果做一个描述。你可能会说“这很管用，使用想象的学生会做得更好。”当然，对聊天来说这就足够了，但作为实践报告或考试问题的答案却不够。“他们做得更好”到底是什么意思？这个演示或测量的精确度如何？由于其他研究者可能想要检查我们所做的研究，因此当我们在心理学研究中定义变量时必须做到精确。那么，对此研究中变量“记忆量”的操作定义可能为：

在呈现的30秒内，将名字和面孔正确匹配的数量。

为了使实验更加标准化和公平，我们必须限定时间，以防出现被试一个星期回来后说“我刚记住！”的情况。

从下面的练习6中你将看到研究者在如何定义变量方面是很挑剔的。原因是他们希望观察的是真实的变量变化而不是使用“粗糙的”证据。你的导师可能（或应该是）对你所下的定义也很挑剔，因为你正在学习的是如何清楚地向别人表达自己的想法。心理学家的交流必须清晰、明了，因为他们希望彼此的研究都是可以重复验证的。

练　习

1. 你如何量化“爱”？
2. 你如何知道某人害羞了？
3. 为什么心理学研究者总是为每件事寻找一个数字呢？
4. 尽力想出一个纯粹的心理学构想。
   提示：对那些可能有某种你无法观察到的心理素质的人，你该如何描述？
5. 为什么心理学研究者（或者老师们）对准确定义一个问题如此小题大做？
6. 在下面的练习中，试着为左边字面粗糙的变量作一个严格的测量定义。在自己想出答案之前先不要看右边我给出的参考答案——不要欺骗自己！

体罚	❑ 父母报告的每周打孩子的次数
	❑ 儿童报告的每周被打的次数
	❑ 对体罚态度量表的测量得分
压力（在工作中）	❑ 雇员报告和压力相关的疾病产生的频率
	❑ 过去6个月内缺勤的次数
	❑ 知觉压力量表的测量得分
三岁儿童的社交性	❑ 与其他儿童说话的时间（以秒为单位）
	❑ 给其他儿童玩具或物品的次数
	❑ 一定时间内目光交流的次数
对罗宾·威廉姆斯的喜欢程度	❑ 被试在量表上的评分（“1表示非常不喜欢”到“7表示非常喜欢”）
	❑ 被试在给定的10个明星中选择威廉姆斯的次数
服从	❑ 被试是否同意给街上需要零钱的人钱
	❑ 问卷调查时，返回问卷的被试人数的百分比
创造性	❑ 解答10个发散思维问题的时间（以秒为单位）
	❑ 两分钟内想出砖头用途的数量

答　案

1. 认真思考你就会发现这是一个很糟糕的任务，我们如何知道什么时候我们或他人在恋爱呢？我认为理解别人更容易，因为我们可以使用心理学量表，里面的题目包括“我禁不住想起某某”等（见6.3节内容）；我们可以观察亲近行为（其实也没那么近！）；记录花在买礼物上的钱；买礼物的次数；买哪类礼物，等等。

2. 大家会觉得这个稍微容易点，害羞的人在做介绍时会犹豫畏缩；当有人靠近时他们会结巴；他们会脸红，等等。
3. 因为他们想知道心理特征或行为是否和其他事情相关或受其他事件影响。为了确定 X 的变化是否由 Y 变化引起，你需要观测当 Y 变化时，X 是否变化。所以即使测量的很粗糙，你还是需要测量。大多数心理学家认为通过这种方式才算是真正的行为科学。然而，定性研究者却不这么认为。（见第 7 章）。
4. 这里通过举些例子来说明：如人类的智力、敏感性、对大自然的关怀、良心、坚持、对迈克尔·杰克逊的态度、偏见、精神等。所有这些都能通过行为观测和提问来评价，并始终可以被我们用作证据。但我们也假设有种受外部世界影响的连续性的东西存在于个体内或者总是同个体联系在一起。个体是如何受影响的，这些影响与什么有关，这才是心理学研究的本质所在。
5. 答案就在练习栏上面的段落里。
6. 见上面练习的右栏。

关键术语

操作性定义（operational definition）	数量化（quantify）
心理学构想（psychological construct）	变量（variable）

2.2 人类——心理学研究的对象

前面我们已经探讨了心理学家使用的测量的种类，我们需要在研究中对测量的定义作严格的规定。现在是时候来看一下心理学中的研究对象，以及在对人类被试做一般性描述时可能会遇到的陷阱了。我们需要意识到的是，当我们使用方法时，所用的方法必须是科学的。科学方法包括一套设计好的程序，这些程序可以用来检验关于世界的一般观点。以生物学家为例，他们并不是对某朵花感兴趣，而是对植物如何留住水分和为什么偏爱酸性土壤感兴趣。同样地，心理学家所感兴趣的是如何把他们从人类样本中获得的结果推广到整个人类——就像尽力从几片叶子的情况推广到整片草坪的生长情况一样。

当一群新生谈论人类行为时，总有一些人略带顽固地作出概括"每个人都是不同的，因此你不可能对人类作一个科学的描述"。事实上，你是可以作出概括的。如果你立了一个警告牌说现在某一路段安装了测速照相机，可以肯定地说，司机们开车一般会比以前慢一些，这样的情况至少会持续一段时间。当然，有些人会保持原来的速度，进一步调查这些人有什么共同特征将会非常有趣。但对于人类的一般行为，我们可以作出相当准确的预测并能观察到相应的变化。

上面最后一句话的意思是，尽管不能事先确定某个特殊个体会产生期望的行为变化，但我们可以对人类总体进行预测。如果我们知道了有些人在作判断时可能会受其他人的判断影响的话，我们就可以设计一个验证性实验。到目前为止，我们还未到达一个确信大部分个体都受到影响的时期。做一个同样粗略的类比，就像生物学家不能预测田里哪一棵植物在施了一种新的肥料之后长得最快，但他们能肯定地声称田里植物整体的生长速度会加快。

样本和总体

这里我们必须认识到的是，心理学研究者们对他们测试的样本人群并不是特别感兴趣。多么无情的观点！我的意思是，就像生物学家一样，心理学家测试样本的目的是为了能对抽取样本的总体进行概括（见图 2.3）。**总体**不必是一个国家或小镇的所有人，而是我们认为我们的理论（正在验证中）所适用的所有人。例如，如果我们正在调查杂志上刊登代表老人观点文章的方式，我们就把所有杂志作为总体。在心理学研究中，无论目标总体（我们对推广感兴趣的群体）是什么，我们总是仅仅检验从总体中抽取的**样本**。

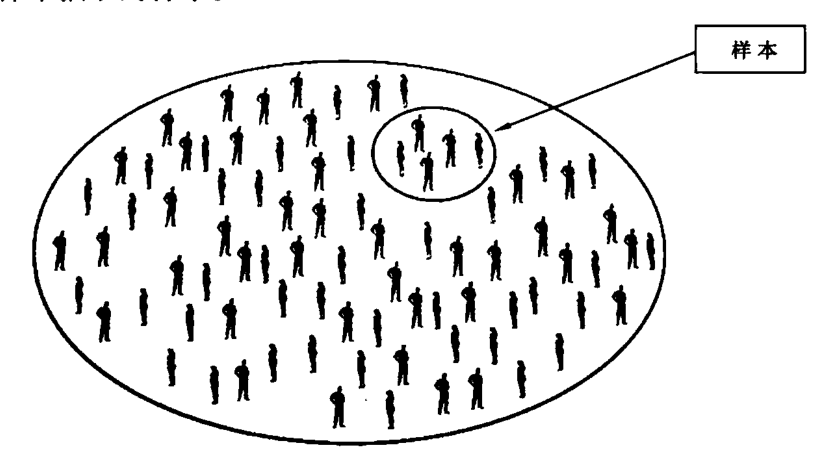


图 2.3　心理学家从总体中抽取小样本进行研究

这就是为什么不论被试是否仅仅是学生、受过高等教育家庭的儿女，还是来自于大城市的儿童等，我们都如此在意我们所检验的样本的原因。实际上，我们目前所担忧的是研究结果的推广。众所周知，推广问题即总体效度问题是研究的核心，本书第 3 章中会更为详细讨论外部效度这个问题。现在来看看图 2.4 中的偏差样本。

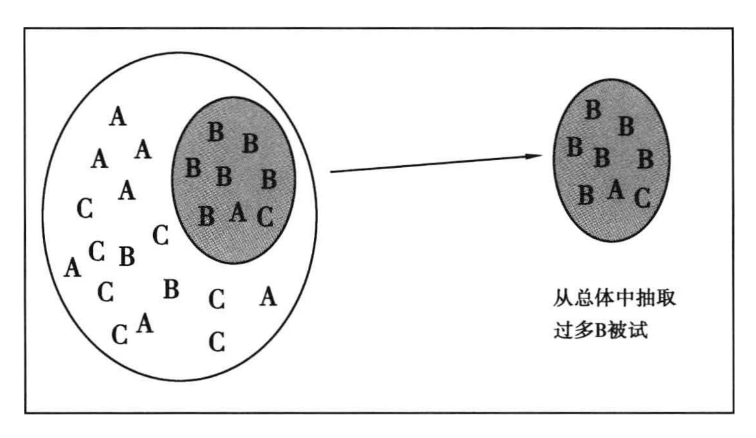


图 2.4　偏差样本

我们需要的样本对于总体而言要尽可能的典型，因为我们希望能将我们的结果推广到总体中。设想我们出于某种原因想要比较男女的开车行为。再设想我们在

上午8:45进行取样观察,并在下午3:15再次观察。那么,我们的样本就包含了比平时更多的后座带着小孩的女性驾驶者,显然,这一点会直接影响这些车辆被驾驶的方式。

当一个样本过分强调某一种类特征时,我们称之为**偏差样本**(biased sample),并认为取样过程有**抽样偏差**(sampling bias)或**选择偏差**(selection bias)。那么,心理学研究的样本为何会出现偏差呢?

心理学研究样本的偏差

非常多的心理学研究是以**志愿者样本**(volunteer sample)为对象开展的。人们不是被强迫走进心理学实验室,而通常是通过在大学心理学系、医疗中心、学校、体育组织张贴公告,非常温和地询问他们是否愿意参加实验(见专栏2.1)。

专栏2.1 心理学研究中自愿者的天性

追溯到1965年,奥拉(Ora)发现,在依赖、不安全感、攻击性、外向性和对社会影响的易感性上,心理学研究中的志愿者比非志愿者得分更高。从那时起,研究者进行了许多的研究。罗思诺和罗森塔尔回顾了之前的研究(Rosnow and Rosenthal,1997),认为相比非志愿者,志愿者的以下特征得到了相关的证据支持,具有"最大""相当"或者"一些"可信度。

最大可信度:良好的教育、高智力、追求认同、社交性

相当可信度:寻求激励、不专横、不顺从、非传统

一些可信度:利他的、自我封闭、适应不良、年轻

因此,有可能心理学研究的一些结果与研究中使用的偏差样本的特性有关(如志愿者)。但不要急于下结论说,所有心理学的研究结果或多或少地令人质疑。正如我们之后会看到的,如果研究设计得非常好,除非研究本身致力于人格变量,否则被试变量的影响会很小。但我们不能过于自满,必须记住,挑选被试时所使用的方法会影响将来结果的解释。

心理学研究中学生被试的普遍性

比志愿者偏差更严重的是,在心理学研究中学生被试占绝对多数。班牙德(Banyard)和亨特(Hunt)查阅了1995和1996年出版的两种英国主要的心理学研究杂志,发现只有29%的研究项目使用非学生成人样本,而其中有几个研究的成人被试是大学的教职员工。在构成这些项目样本的18 635人中,学生占67%,其中心理学专业学生占相当大的比例。

这一现象并不太令人吃惊,这是因为心理学研究者邀请最接近他们的大量学生参与他们的研究更为容易。与之相一致的是,塞尔斯在大约14年前就发现在美国学生被试呈现相似的比例(Sears,1986)。这对于我们这些大西洋对岸的人来说并不惊奇,因为在美国心理学学生被要求参加一些研究项目或应付额外的书面作业有着很长的传统,因此学生只是名义上的志愿者。现在英国大学也采取了相同的体系,例如,学生必须参与实验才能获得学分,获得他们自己第三年项目的被试,或者换取印刷服务,等等。

2.3 抽 样

代表性样本

理想上(即使在实践中几乎不会发生),我们需要的样本能代表一般的总体,或者代表研究所关注的总体(如男人、妇女、八岁儿童、护士等)。实际操作时,我们要尽力做的就是尽可能多地排除明显的抽样偏差。也就是,我们尽力确保总体中的某类人不比其他人有更多机会进入我们的样本。被试类型会随着研究项目的类型而变化。如果是视知觉实验,我们就不想要太多当地摄影俱乐部的成员;如果是儿童和逻辑思维的研究,我们就要避免高智商协会(MENSA)最新招募的儿童。

"随机的"并不意味着只是"随意的"。严格地说,一个随机选择序列意味着无法从前面的序列预测出后面的序列。当人们认为他们在进行随意选择时,可能还是有一个潜在的模式,只是他们过于高兴而没有意识到罢了。对蝴蝶来说不是这样。大自然赋予蝴蝶能够在飞行时进行无限的随机翻转的能力,以确保任何猎食者都无法作出预测。

随机样本

你认为下面哪一个程序会使你使用随机样本进行研究?在每种情况下,我们都假设目标总体已经确定了(如在第1项中为一般公众)

1. 只选择一侧街道上的人。
2. 在公司雇员表上,每隔10个选择1个。
3. 在所有伍布雷学院学生的名单上用针刺孔。
4. 从包含伍布雷学院所有学生名字的帽子里选择纸条,让选中的人填写性别行为问卷。
5. 在诺莱斋学院食堂里随意地接近人。

(1分钟以后回答)

简单的随机样本定义的其中一部分是,目标总体的每个成员都有相同的机会被选中,我们应该关注这个标准(定义的另一部分是任何组合都有可能出现)。这一原则是建立在运气的基础上。学生在他们的实践报告里写到:

"我们从总体里选择了一个随机样本……"

不要尝试在你的报告里写这样的话。不只是对你而言,对几乎所有的心理学研究项目而言,可以肯定这不可能是真的。想想这意味着什么。不管你用何种方法收集被试(接近大众人群、邀请朋友、在食堂提出邀请,等等),病人、犯人、轮班工人、石油钻井工人等所有的人都有相等的机会进入你的样本。当然不可能!

那么,上面练习的答案是"没有一个!"。在第1、5题中,我们不会问那些看来不容易接近的人,而且无论如何我们只是从在街上或食堂里的人中取样。第2题明显不符合可能性均等的随机选择标准。第3题的程序可能会偏向页码的中央。在第

4 题最初选择了相当的随机样本，但极有可能同意参加研究的被试数，会因问卷的特性而大幅度减少。在所有诸如此类的例子里，某些潜在的被试无法获得或拒绝参与，这本身就是有趣的偏差性因素。

你能从有限的总体里挑选随机样本，如你课堂上的所有学生或你大学里的所有学生，但你仍然不能说一个随机样本参加了你的研究，因为几乎可以肯定的是，你选择的许多人不能参加研究。

随机分配到实验条件中

你必须仔细区分从总体中挑选随机样本（几乎不可行）和将被试**随机分配**（random allocation）到各个实验条件中（非常普通和相当容易实行）。在一个实验中我们需要分配被试，比如说，把 20 个被试随机分配到两种实验条件中。也就是，每个被试都有相同的机会被分配到任何一个实验条件中去。我们需要的是没有选择偏差。为此，可以使用专栏 2.2 中的方式，非常简单地从 20 个人中选出 10 名被试（见图 2.6）。你也可以抛硬币，如果背面朝上，那么第一个人进入实验条件 1，第二个人进入实验条件 2 等——我们马上会对系统随机抽样进行描述。你甚至能为每个被试抛硬币，但如果在另一条件下已经使用了抛掷硬币的方法，那么你就必须把最后几个被试分配在同一个实验条件中——这并不是一个大问题，但仍不够理想。随机分配不会把被试分成极为相等的两组，但对大部分研究而言，我们可以假定我们已经尽力排除了所有抽样偏差的影响。

专栏 2.2　如何随机取样

如果你需要从 100 人中随机选择 20 人，有几种方法可以做到。首先，你给每一个人从 1 到 100 中的一个号码。接着你需要一种方式从中随机选择 20 人。

产生随机数字：使用计算机或附录 2 的表 1 产生 20 个随机数字。使用随机表时，你任选一个起点，按同一方向继续，在移动的同时记录每一个数字。根据产生的数字挑选 20 人。

抽签法：你可以在尺寸相同的纸片上写下每一个数字，并将它们放入箱子里，接着用力摇晃，然后挑选 20 个数字，由此你可以尽可能地做到随机。然而，摇晃是最重要的。如果只是把数字放入箱子里，就会因为数字排序的方式，明显地造成偏差。

随机顺序的其他用途

实验中有时需要随机排列一组刺激。刺激可能是一组需要学习、回忆的数字或需要解决的字谜。你可能也想随机呈现一系列实验测试——例如，用左手或右手使杠杆平衡。每个被试每只手完成 10 次测试，我们想随机安排左右手实验次序。像上面描述的，在任何一个例子中，你赋予每个刺激（例如词汇）或测试一个随机数字，然后把随机数字按顺序排好。这样与随机数字相对应的词汇或测试就排列成随机顺序。这个过程有时被称为刺激或测试的**随机化**（randomisation）。

系统抽样

另一种从较大的总体中挑选被试的方法是**系统抽样**（systematic sampling）程序。

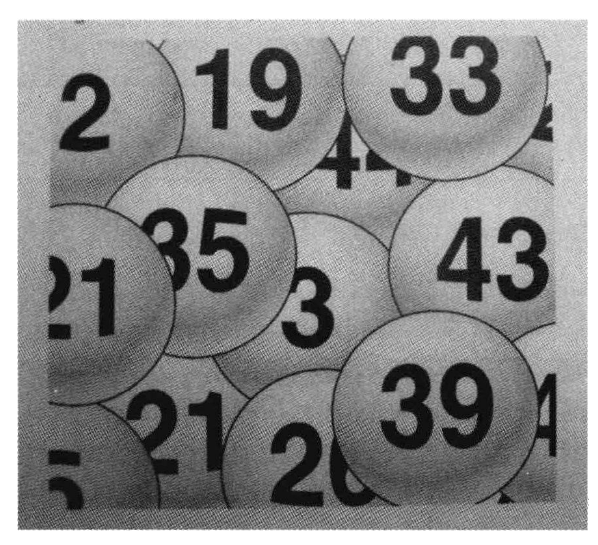


图 2.5 摇动乐透彩球过程中的随机选择

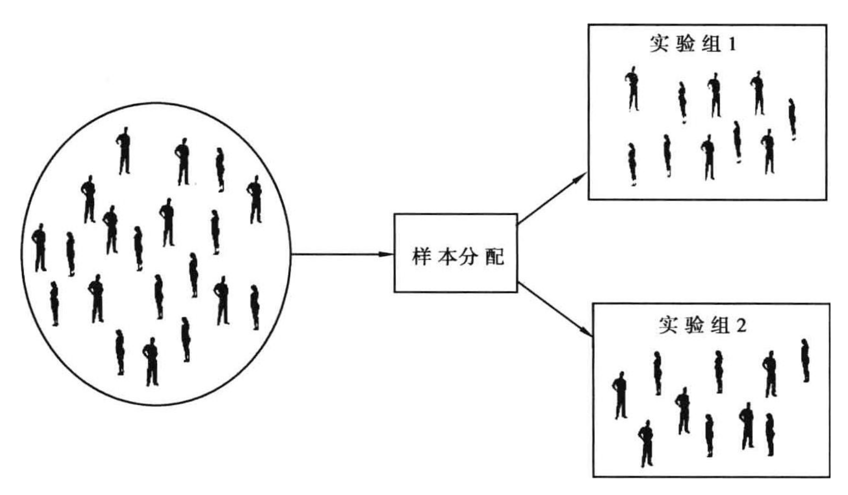


图 2.6 随机分配到实验条件中会产生等组(但不完美)

为了从 100 人中选择 10 人,我们可以简单地从 100 人的名单中每隔 10 人挑选一名。如果我们想要每个人都有相同的机会被选中,我们一开始就从 1 到 10 之间随机选择一个数字,用这个数字作为起点。如果我们的随机数字是 7,那么就挑选第 7 个人、第 17 个人、第 27 个人,以此类推。

分层抽样和整群抽样

假设你不信任随机抽样能够从学校总体中抽到能够代表所有课程或者科系的样本,那么就从学科分布的角度进行抽样。简而言之,试想一下大学有四个主要的科系:艺术、自然科学、社会科学和商务研究。假设每个科系的学生比例如图 2.7 上端所示。我们可以在每个科系内随机抽样,并确保在最终的**分层样本**(stratified sample)中每个科系的学生比例与在总体中的比例相一致。注意在这个样本中,每个人有相同的机会被选择而且任何最终的组合都是有可能的。

如何选择相关的层取决于研究项目的本质。如果调查对失业的态度,我们希望

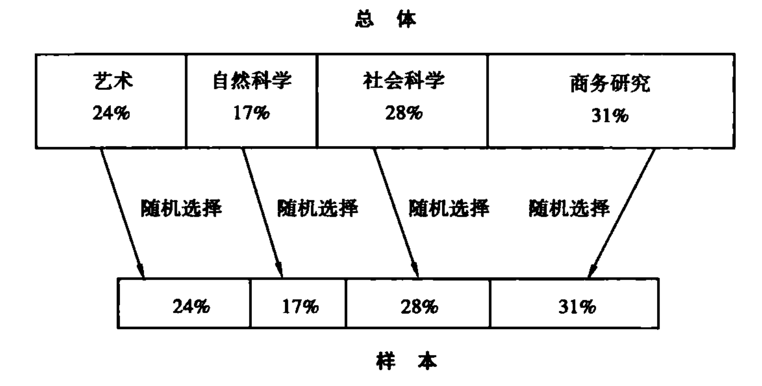


图 2.7 作为分层样本的大学各部门的比例

能按比例选取雇员、失业者、个体经营者、家庭雇工和囚犯。例如，我们不能只选择50 人，并使之在如下的方面代表当地的总体：性别、健康状况、教育水平、宗教、投票意向和种族。我们必须选择最相关的层或选择相当大的样本。

分层样本的一种形式是**定额样本**（quota sample），不过这只是一种很不成熟的方法，例如只有当街头访谈者已经被告知要询问多少定额的女性和男性时才会使用；这种取样方法并不科学。

非随机样本

机会样本或方便样本

通常，特别是在学生的实践作业中，我们从能够轻易掌控的人群中招募研究被试：同学、朋友、亲戚、导师允许我们从中招募被试的一个班（在课程允许情况下），或只是选择我们在大学食堂里可以接近的人。

很明显，这种样本被称为**机会样本**（opportunity sample）或**方便样本**（convenience sample）。当然，一个样本被称为机会样本并不会告诉我们被试是如何被招募的，因此在实践报告中从来不要尝试只使用这个术语。近来我在和卡拉·弗拉纳根（Cara Flanagan）讨论研究方法，她是一位资深的 A 级课程主考人员。谈论中，我提出短语“对正式术语的敬仰”，以描述导师和学生的一种带有高焦虑的需要，在这种状态下他们为每件事寻找一个名称，如果找到了，就感觉现在所有的问题都解释清楚了。于是，我想到了这种术语（机会样本）。你会感到所有的抽样方法必须有一个名称，如果有了名称，读者自然就理解了有关样本的一切。那么，无论何时我的学生脱口而出“被试是机会样本”，他们就会收到这样的评语：“他们是如何被招募的？”我们想知道他们的动机是如何产生的、他们是如何被接近的、当时他们在做什么，等等，因为这些因素都可能影响结果。如果他们是自愿者，例如之前提到的奥拉、罗思诺和罗森塔尔的例子，我们知道他们会形成一个不同于正常情况的样本。因此，不要说你抽取了机会样本。请描述被试是如何被挑选的。

自我选择样本

我们之前看到，相比总体平均分，志愿者在某些特性上的得分或高或低。自愿

者选择参加研究,因此我们称之为**自我选择样本**(self-selecting sample)。换言之,被试自己选择参加研究,而这可能是被试有意的也可能是无意的。例如,设想一下,我们感兴趣的是,如果街头卖艺人穿得好些,人们是否会给他更多的钱。因此我们让卖艺人的表演分两阶段进行———一会儿穿得漂亮,一会儿穿得破旧。我们的样本只是那些给钱的人,这些人显然是自我挑选的。我们不知道那些没给钱的人是否也注意到了衣服。

专门小组、焦点团体和滚雪球抽样

专门小组、焦点团体和滚雪球抽样都是定性研究中广泛使用的收集数据的方法(见第9章)。**专门小组**(panels)是有良好分层和代表性的群体,在相当一段时间内被邀请参加几个研究活动,由于他们有良好的平衡性,因此非常有价值。**焦点团体**(focus group)倾向于因为特定的目标而会面,常常可能是对一个特定问题的讨论,如对儿童的体罚或健康饮食态度。**滚雪球抽样**(snowball sampling),例如研究者会见一个关键人物(如社团领袖),进行一次面谈,接着和其他地位高的人谈谈看法。一次接触产生更多接触,这样样本越滚越大。

大样本还是小样本

如果你有时间和资源,样本越大越好。小样本更可能产生抽样偏差。举一个简单的例子,设想我们从包含5个穆斯林、5个基督教徒、5个锡克教徒、5个印度教徒的群体中选择5个人,那么可能只选中一个印度教徒。但是,如果是10人样本,这种情况就不可能发生。因此,样本越大,抽样偏差的可能性越小。

与之相反,这并不意味着总是需要大样本或大样本是必不可少的。不需要大样本是因为总体中的一个小比例的样本就将给你相当好的估计。在竞选期间查阅一下报纸,你会发现报道的民意调查会对最后的竞选结果产生极好的估计。而这是通过在超过4 000万的总体中只挑选一小部分——1 500人产生的。

反对在实验工作中采用大样本的观点认为,如果你需要许多人来证明很小的效果,可能意味着你需要重新设计实验。例如,如果你的实验需要大样本来证明咖啡因提高反应时,那可能是你使用的咖啡因剂量太小,或者咖啡因只对某类人有影响。基于这些考虑,你需要重新设计实验,以尽力展现更明显的效果,而不只是让你的研究陷入需要几百人的困境中。

对于现场研究,特别是问卷调查,大样本会更好,因为在实际操作中有太多无法控制的无关变量,所以即使是微弱的效果我们也希望能够表现出来。那些基于科学的报刊文章通常称之为“效果小但有显著趋势”。这种趋势总是可以通过多种方式得到解释,它们的存在非常有价值。

练 习

1. 一位心理学家在大学报刊上刊登广告招募学生，希望他们参加有关酒精消费对食欲的影响的实验。你认为什么原因导致被吸引来参加实验的样本不是随机的？
2. 在撒非科斯郡，下面哪一个方法能产生真正的心理学学生的随机样本？
   ①随机挑选一个学校，要求心理学学生自愿参加。
   ②从每个学校中随机挑选学生，他们的名字分别以字母表中的每一个不同字母开头。
   ③把每个学校的所有心理学学生的名字放入帽子里，摇晃之后抽取样本。
3. 心理学老师进行实验，她把班里前半部分学生分配到一种实验条件，并教这些学生特定的问题解决方法。而班级后半部分的学生作为控制组。实验组成绩更好。那么，除了指导语的不同，实验程序中还有什么缺陷导致了学生间可观察到的差异？
4. 一个学生写到她为研究项目收集了一个“机会样本”。这就是她的导师（或者审查者）想知道的一切关于被试的事情吗？
5. 我们为什么把研究项目中的被试称为“样本”？

答 案

1. 参与的学生会是自愿者；不包括滴酒不沾的人；样本偏向于阅读报纸的人，或者是由读报的人转述的。
2. 只有方法③包含（不随机的）自愿者。方法②不能给学生同等的机会（如果你名字以“z”开头，就比以“t”开头有更多的选中机会）。甚至方法③如果不是所有学生都同意参加，也不能产生最终的随机样本。
3. 她没有随机分配被试到实验条件。也许班里前半部分学生对实验更感兴趣，因为他们受到了较好的教育，并且他们更擅长用问题解决任务。
4. 不是。她应当准确解释被试是如何招募的。
5. 因为他们只是一个小群组。我们不能一次验证整个总体。我们可以认为，从样本中获得的效果在样本所在总体中很可能发生。样本要代表总体，我们才能把样本结果推广到总体。

关键术语

偏差样本（biased sample）
方便样本（convenience sample）
焦点团体（focus group）
概括化（generalization）
机会样本（opportunity sample）
专门小组（pannel）
总体（population）
定额样本（quota sample）
随机分配（random allocation）
随机化（randomisation）
随机样本（random sample）
样本（sample）
抽样偏差（sampling bias）
选择偏差（selection bias）
自我选择样本（self-selecting sample）
滚雪球样本（snowball sample）
分层样本（stratified sample）
系统抽样（systematic sample）

3

实验方法

本章内容

- ❑ 描述了真实验的结构,其中包括变量的控制和操纵,使读者可以看到自变量对因变量的影响。
- ❑ 未控制变量或无关变量会成为干扰变量,掩饰真实的实验效果或令实验缺乏效果,产生虚假解释。
- ❑ 讨论几种实验设计:独立样本、重复测量、配对。
- ❑ 实验室实验和现场实验或研究之间的差异,以及它们之间的优缺点。
- ❑ 有关内部效度、外部效度的更多话题,以及对生态效度的误解。
- ❑ 最后讨论实验和其他研究的多种偏差来源:研究设计、被试、研究者期望和影响。

3.1 可替换性解释

我们已经在第1章中考虑了多种可能的方法来收集证据，以证明“炎热引起攻击”的观点。心理学家常常从人群中直接收集证据，但也借助社会统计学来**间接**收集证据。例如，如果炎热引起攻击，那么随着温度的增加，对于任何特定区域的犯罪记录，我们期望看到何种统计？假设我们期望看到，犯罪随着温度的增加而增加，即温暖的季节比寒冷的季节会引起更多的攻击和伤害等。

> 安德森(Anderson,1987)在北美的几个大城市中收集了这类犯罪记录，结果表明，温度越高的年份，暴力犯罪率就越高。调查结果同时发现，一年中某一季度的温度越高，暴力犯罪越多，甚至城市的温度越高暴力犯罪也越多。
>
> 你认为这种证据会对“炎热引起攻击”的假设提供了强有力的支持吗？除了“炎热”，你还能想出导致暴力犯罪增加的其他原因吗？

摆在我们面前的难题是，随着温度的增加，其他因素也在增加。例如，天气越温暖，在公共场所的人就越多，排队购买冰激凌的人就越多，交通堵塞的时间也就越长，等等。可能根本不是炎热而是其他因素引起了攻击(见图3.1)。这种**非实验设计**(non-experimental design)的问题在于，我们无法排除其他因素，因为这些因素在日常生活中总是起作用的。

我们可以通过收集数据来支持假设，但在诸如“炎热引起攻击”的许多研究中我们无法排除**竞争性解释**(competing explanation)，即我们无法明确事件和人类行为之间是否存在清晰的因果关系。我们根本不能确定是热引起了观测到的人类攻击行为的增加。在这种设计中可能有其他不能控制的变量。而在后面将要看到的真实验设计中，就能够对这些额外变量加以控制。

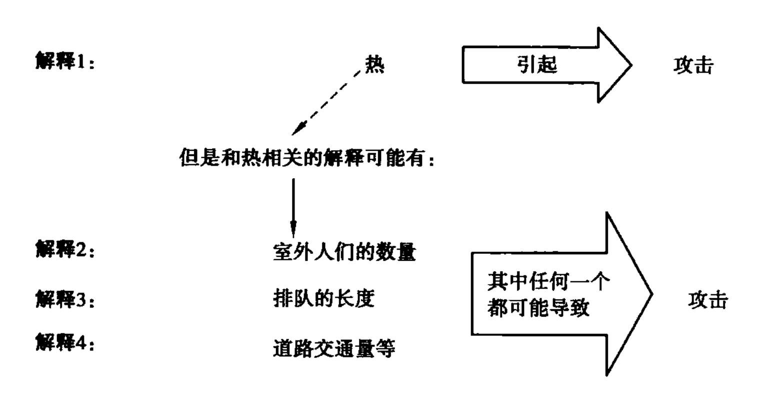


图3.1 热-攻击联系的可替换性解释

再来看另一种包含更多控制的设计类型。

> 肯瑞克和麦克法雷那(Kenrick and Farlane,1986)在亚利桑那州 Arizona 的菲尼克斯 Phoenix 观测了男女司机们在等待绿灯时的按喇叭行为。当时,亚利桑那州气温在华氏 88 度到 116 度之间(大约为摄氏 31 ~47 ℃)。他们发现,温度与按喇叭行为之间有直接关系,随着温度的升高,按喇叭的行为也在增加。实际上,对那些将窗户摇下的司机来说,这种关系更强——假设因为车内空调未开而导致摇下窗户的司机会更不舒适。

这种设计包含更多的控制,因为研究者能直接观测到司机做了什么,而且攻击被限制为一个简单而清晰的抽样行为——按喇叭。但研究者无法控制任何一天的温度,也无法控制测试的人们,因为他们可能在那天不得不开车出门。设想一下,或许攻击性高的司机倾向于热天出车,而低攻击性司机倾向于热天不出车。我们再次看到,在热和攻击这一明显的因果关系中存在着其他解释。

问题仍然在于研究者没有进行真实验研究。现在让我们从一个简单的例子来看什么是真实验。

> 假如你正与一些老熟人一起吃饭,你注意到窗台上的华丽花卉。你那位神秘的密友告诉你,与电视机上的萎靡不振的花儿相比,窗台上的花儿之所以会变得如此华丽,是因为他们每天与花儿交谈,鼓励花儿成长。你看到窗台上的花儿吸收更多阳光,栽在更大的花盆里。所有这些因素都可能导致窗台上和电视机上的花儿出现差异。当你试图反驳这种谈话能促进植物生长的理论时,激烈的争论开始了。
>
> 你怎样设计一个公平的实验来表明与植物对话对它们的成长到底是否有影响?

与第 1 章类似,我希望通过下面的练习来说明,未受过心理学训练的人也能很容易弄清楚真实验的构成。希望你的设计像这样:

> 挑选同一种植物的两个样本,置于相同的容器内(含有相等的土壤和营养比例),放在两个地方(拥有相同的阳光和热)。随着植物生长,与一组谈话,不与另一组谈话。12 周后测量它们的成长(如高度、宽度、色彩深度等)。

当然 ,我们还没有定义什么是更好的发展,因此目前只能进行相当有限的测量——植物高度。

1. 在这个设计中包含了实验的所有基本要素。
2. 实验者控制和操纵了**自变量**(independent variable)。
3. 其他变量或被排除或被恒定。
4. 实验者测量**因变量**(dependent variable)的所有变化。

因此,在实验中包括:

> 自变量:植物处理。
>
> 因变量:12 周后植物的高度。

自变量的水平

自变量是有**水平**之分的:上例中,我们不说自变量是“谈话”或“不谈话”。要养

成习惯去寻找自变量的一般主题，如“植物处理”，然后定义自变量的水平——这里为“谈话”和“不谈话”。当我们处理自变量时，这一点就变得比较明显了，比如说在咖啡因研究中，水平可能是5毫克、10毫克、30毫克等。注意：实验并不总是或仅仅通常只包含两个水平。本书只是从这种简单的情况开始而已，事实上实验可能包括5个以上的条件。一般设计包括三个条件——实验条件、控制条件、安慰剂条件。

在这个实验中“不谈话”可以被称为控制条件，因为我们需要比较在实验条件下实验组里的植物，它们没有接受任何的处理，用于基准线测量。我们所决定的因变量就是12周后所测试的植物高度。

3.2 实验包括控制条件

在前面的暴力犯罪和按喇叭行为的统计的例子中，我们注意到其他解释的可能性和多样性。这些**并不是**实验的例子，因为研究者没有操作变量。上面的每一个设计都有几个变量没有仔细控制，但在实验里我们要对变量进行**控制**。可以说，只有自变量是唯一变化的，当观测到因变量随之发生相应的变化时，两个变量之间就存在着因果关系，这正是我们通过操作自变量X来隔离因变量Y的原因。执行实验的基本部分就是尽力控制或排除**无关变量**(extraneous variable)——无关变量可能会与自变量或因变量产生相互作用，从而混淆了是否是自变量直接影响了因变量的问题。**干扰变量**(confounding variable)是一种重要的无关变量，下一步我们将处理这一问题。

练 习

1. 一个实验的基本特征是什么？
2. 什么是自变量的水平？
3. 对于一个显著效果更可能存在其他解释的是实验还是非实验？

答 案

1. 一个实验包括至少两个自变量水平的操纵，以观测和记录因变量的任何结果变化。为了排除所有的其他解释，其他变量要么保持恒定，要么被以某种方式排除。
2. 自变量的水平常常是实验处理的条件，例如0毫克、20毫克、50毫克的咖啡因，或冷热条件。当在正常条件没有发生变化的情况下执行实验任务时，实验条件就被称为控制条件——例如0毫克的咖啡因或正常温度。
3. 显著效果的其他解释总是可能存在的，但在非实验里，由于变量缺乏良好控制，常常存在更多其他解释的可能。

关键术语

控制条件/组(control condition/group)	无关变量(extraneous variable)
因变量(dependent variable,DV)	自变量(independent variable,IV)
实验(experiment)	自变量水平(levels of the IV)
实验条件(experimental condition)	处理(treatment)

实验与干扰变量

心理学研究的新手习惯把所有研究都称为"实验"。从现在起,请确保自己不再属于新手这类人。实验是一种特殊类型的研究设计,必须具备某些关键特征才能称之为实验。当然,所有研究都不可能是完美的。在上面的例子中,应该注意到为了和植物说话,当你靠近它们的时候就不得不进行呼吸。所以,也许是呼吸引起了植物生长,而不是谈话。有些变量随自变量而变化,并可能为因变量的变化提供其他解释,这种变量称为干扰变量(见图3.2)。

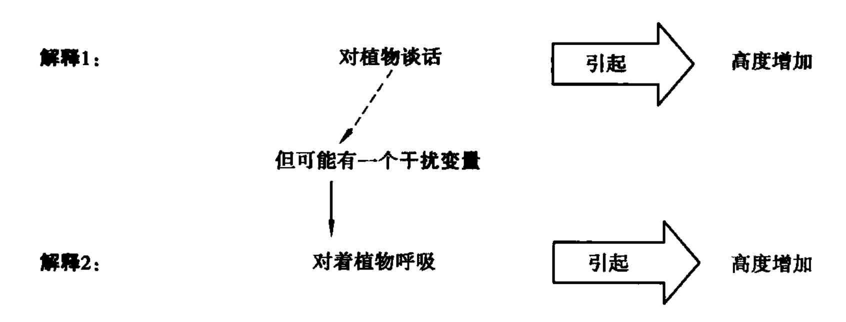


图3.2 植物实验中的干扰变量

科学活动的一个重要部分,就是寻找研究中似乎可以支持研究者理论的可能干扰变量,然后进一步设计实验,或者排除干扰变量,或者**使用**干扰变量进行实验。例如,在植物研究中,进一步研究可能包括保持一定距离与植物谈话,或者使用录音机(排除干扰变量),或者**仅仅**对着实验条件下的植物进行呼吸(**使用**干扰变量进行实验)。

现在来看看研究者是如何利用真实验来验证"炎热引起攻击"这个假设的。

> 鲁尔、泰勒和多贝斯(Rule, Taylor and Dobbs,1987)要求被试在21 ℃、33 ℃等两种不同温度条件下补充完成结局模糊的故事,故事的结尾部分可能会出现攻击性结果。与较凉爽条件下完成的故事相比,较热条件下完成的故事包含了更多明显的攻击内容和负性情绪。

当然,这与现实世界中的真正攻击行为不同。但所有的科学实验都拥有所谓"人为性"的特征。他们只抽取存在疑问的变量,尽量排除其他无关变量,并尽力证明研究的确存在着某种影响。在这个例子中,似乎只有温度变化导致故事结尾内容的差异。但实际上至少还存在着其他的可能性(啊,但是……!),这些我们在后边会论述到。但你是否能从中看到关于差异性的其他解释。

保持控制——标准化程序

我们已经了解到体现实验优势的一个关键因素在于,只要控制了实验中所有其他无关变量,研究便有可能表明因果关系。正因为如此,实验使用标准化程序来收集数据。研究者如果对一组被试非常友好而对另一组被试却有一点脾气古怪,这对实验是毫无用处的。如果因为被试紧张就向他们透露更多执行任务的信息是非常糊涂的行为。我们需要每个被试都准确地在相同条件下执行任务,使除了自变量外

所有的因素都保持相同。研究者使用**标准化指令**，意味着给每个被试都有相同的指导语，确保任何人都不可能获得更多的信息。

重复研究——清晰、完整和准确的报告的需要

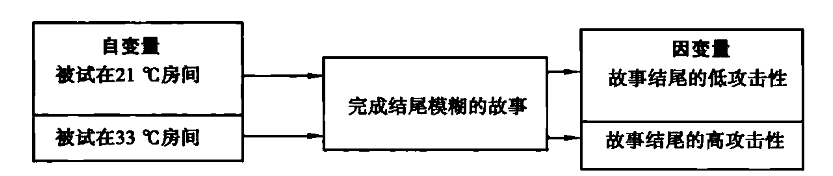
假设你读到这样一个令人惊奇的效应。你听说心理学家要求被试裹上昨晚某人睡觉时的毯子，并且通过四张图片来选择当晚那个人做梦的情境。他们的表现远远好于随机水平。这是一个虚构的研究，但如果你读后，你肯定想准确知道研究是如何进行的，被试是谁、指导语是什么、使用了何种设备，你需要一个清晰准确的研究报告。这就是为什么你的导师鼓励你在实践调查的科学报告中确保完整性和准确性的原因。他们不是虐待狂，而是帮助你形成传达信息的技巧，无论在学习过程还是在将来都对你有好处，在任何条件下你都要给读者翔实和精确的信息，让他们知道你做了什么。

你可能想准确地知道研究是如何进行的，因为你可能想重复这个研究，即为了看看你是否能获得相同的效果而进行重复研究。如果一个效应获得了重复，我们会更相信它是真实的，即有更高的效度。不是所有的"重复"都能准确地复制原始研究。比如，尽管有些愚蠢，但我们可能会猜疑，效应发生是由于正确的图片总是呈现在左上角。我们更偏爱左上角，因为它们首先映入眼帘。重复包括了原始研究的所有重要特征，除了在一个方面进行修改——每次随机化正确图片的位置。我们试图用这样的方法排除可能存在的干扰变量。

3.3 实验设计类型

独立样本设计

回到前面鲁尔等人(Rule et al.,1987)的研究中。他们使用了下面的实验设计类型：



为什么他们不要求相同的被试首先在冷房间完成故事，然后再在热房间完成故事？如果是这样，想想实验会发生什么。被试可能会确切地知道实验是关于什么的，并且会像他们在第一个条件下一样完成相同的故事。这就没有意义，不是吗？我们在实验中使用这种操纵，两种条件或自变量水平就应该使用两组不同的被试，这点非常重要。

不幸的是，我们第一次遇到了心理学研究方法所具有的一个共同特征。多年来，心理学家一直从生物学、物理学和社会学等其他学科的科学家身上汲取养分，结果把那些有着相同意思的不同术语引进了心理学研究。对于通称为**独立样本设计**(因为两组被试样本相互独立)的上述设计，我们也使用相同的术语：

独立分组
独立被试
独立测量
组间

> 所有术语都指向一种简单的设计，在这种设计中，自变量的每一个水平都用不同的被试来测试。他们属于**独立设计**（unrelated design），因为自变量一个水平的得分与另一个水平无关。

非等组——被试变量问题

我们已经遇到这类设计的共同问题。对于肯瑞克（Kenrick）和麦克法雷那（MacFarlane）的按喇叭研究，我们可以看到什么？也许更多攻击性高的司机就是在更热的天气里开车而不是在更凉爽的天气里开车。如果在独立样本设计中，当更多的某类被试卷入其中一个实验条件时，我们会遇到严重的干扰变量。如果在鲁尔的故事完成研究中，热条件下的被试更有攻击性，那么就可能是这种**被试变量**（participant variable）导致故事结尾更有攻击性，而不是温度。被试变量指人们之间的差异，它可能是所有观测效果的真正原因。

随机分配给不同的实验条件

为了处理非等组可能存在的问题，我们肯定不能让每一组被试都有相同的攻击水平，尽管理想上应该这样。但是我们要尽可能地接近理想。分配被试到不同实验条件的最普通方法是随机分配。即被试拥有相同的机会被选择进入任何一个实验条件。请不要像学生们经常做的那样混淆随机分配和随机抽样。

非等组问题的其他解决方法

除了随机分配，我们可以通过**前测**（pre-testing）确保分组是相同的。我们使用攻击性量表（见第6章），取两个最高分把这一对分配到每一个实验条件，以此类推处理所有分组。使用这种方法，最后获得的两组就会大致相同，但必须确保研究前测量表是有效的。我们也可以用类似的方法分配被试，平衡性别、教育背景、年龄等。

重复测量设计

非等组问题彻底的解决方法是让各组相同。即让每个被试测试所有的实验条件。从上面的例子中我们可以看到，这种方法并不现实，因为被试参与一个实验条件会严重影响他们在另一条件下的表现和认知。在鲁尔等人的研究中，被试会准确知道发生了什么，进而调整自己的表现以符合实验的要求。他们也不得不两次完成相同的故事，这是非常不现实的。

当参与一个条件不会严重影响另一条件时，我们就使用**重复测量设计**（repeated measures design）。在这种设计里，每个被试在每种实验条件下（如自变量的每一个水平）都被测试。图3.3展示了一个实验，要求被试执行迷津任务（把球从迷津一

端运到另一端)，既包括独自完成，也包括在观众面前完成，来验证“社会抑制”假设——观众的存在会影响这类任务的表现。

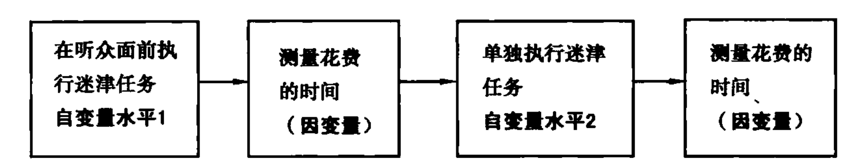


图3.3 相关测量实验

实验中自变量是观众(存在或不存在)，因变量是成功完成迷津任务。

重复测量设计的优点

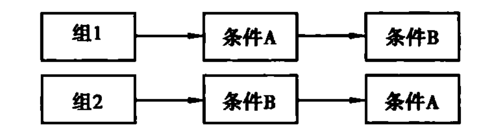
重复测量设计的明显优点是，我们不会遇到非等组问题。每一个被试都参与两种条件。因此我们只需检查每个被试的成绩差异。在某些情况下，被试之间差异可能会与实验结果有关。例如，有些被试已对任务进行了很好的练习，测试成绩可能只有很小的提高，但其他人对任务很陌生，则会产生较大的差异。但是在大多数情况下，不同条件下因变量的差异仍会出现。因此，重要的是不要下意识地批评重复测量设计研究“存在被试差异”。因为我们**只能**看到被试**内**的差异，被试间的差异却被极大地忽略了。最后这一点可以用该设计的另一名称——**组内设计**(within-groups design)来更明确地解释。

尽管不太重要，但与独立样本设计相比，重复测量设计的另一个优点是在两种条件的实验中利用相同被试，可以获得**双倍**的结果，每种被试在每种条件下都提供一个分数，而独立样本设计的被试只提供单一条件下的一个分数。在三种条件的实验中，我们获得三倍量的分数，换言之，在相等的独立样本设计中，我们需要三倍量的被试，以此类推。

顺序效应——重复测量的严重问题

在前面的社会抑制实验里，我们遇到了相当突出的问题。如果所有的被试首先在观众面前执行任务，然后单独执行的话，在第二种条件下他们成绩会有所提高，因为他们在任务里获得了更多的练习。这就是**顺序效应**(order effect)，这是一个经典的干扰变量，但能用多种方法加以处理。

1. **平衡**(counterbalancing)：如果练习只是影响第二种条件，我们就确保一半被试先参与第二种条件。因此，我们可以让一半被试按照顺序条件 A—B 来测试，另一半被试则按照倒序条件 B—A 来测试，模式如下：



这就是所谓的平衡。

重要提示1：平衡不是去除顺序效应，只是把顺序效应分散到两个条件里，以中和顺序效应。

重要提示 2:由于上图存在两个确定的组,因而不要单纯地认为它属于独立样本设计。重要的问题是:每个被试在每种条件下都只贡献一个分数吗?答案"是"。之所以存在两组,只是因为一组按照 A—B 顺序,而另一组按照 B—A 顺序。

2. **预先训练**:训练被试进行迷津任务,使他们获得很好的成绩,直至额外训练不再对成绩有影响。
3. **随机化刺激**(randomization of stimuli):如果自变量每个水平由几个测试条目构成,我们就可以同时在两个水平上进行测试。例如,要测试 10 个具体词和 10 个抽象词之间的回忆差异,不是先测试具体词,再测试抽象词,而是把打乱了顺序的 20 个词同时呈现。之后,再分别计算具体词和抽象词的回忆数量,它们构成了我们所说的两种条件。

配对设计

从前面可以看出,独立样本设计的问题在于,两个实验条件下的两组被试在某些能力或其他心理特征方面并不相等。在许多研究中,我们不能使用重复测量设计来排除这个问题。如果研究涉及诸如对男与女、外向与内向或出租司机与小型出租车司机进行比较等问题时,显然,在每一种条件下,我们都不得不使用不同的被试,因为这里的"条件"就是指被试差异。

有时我们想给予一组人某种训练程序,然后把他们与控制组(除了没有训练内容,其他条件都相同)比较。这里我们也无法使用重复测量。我们能够做到的是把训练条件下的被试与控制条件下的被试相**匹配**。凡奥梅特(Felmet,1998)曾让一组注意缺失多动症儿童进行 8 周的空手道训练,另一组相同症状的儿童在相同的时段里却不进行空手道训练。儿童在年龄和父母教育水平上进行了匹配。结果似乎表明,空手道训练组儿童在某些需要持续注意的任务上表现更好。在这个实验中,自变量是空手道训练,有几个因变量对多动症儿童进行测量。

在这种配对设计里,儿童首先严格匹配成对,然后每对儿童中的一个被随机分配到一个实验条件,另一个进入控制条件。这种设计的优点是,每种条件下都有不同的被试(并且不存在顺序效应或被试猜测实验目的的问题),但我们知道我们能得到等组而且能比较配对被试的分数,就像他们在重复测量设计中的测试的同一个人一样。

相关设计

重复测量和配对设计都称为**相关设计**(related design),因为在这两个设计中我们最终在分析时都获得了分数对。当我们在第 9 章进行统计显著性检验时,这是一个重要概念。表 3.1 提供了每个设计的数据组之间的差异,字母或带数字的字母代表不同的被试。

表 3.1　独立和相关实验设计的数据安排

独立设计		相关设计			
独立样本		重复测量		配对	
条件 1	条件 2	条件 1	条件 2	条件 1	条件 2
A	H	A	A	A1	A2
B	I	B	B	B1	B2
C	J	C	C	C1	C2
D	K	D	D	D1	D2
E	L	E	E	E1	E2
F		F	F	F1	F2
G					

表 3.2　不同实验设计的优缺点

设　计	优　点	缺　点
独立样本	无顺序效应 被试不知道其他条件，不能猜测假设	非等组，在关键变量上存在被试差异 不经济，只能从每个被试那里获得一个结果
重复测量	在能力等方面的个体差异不是很重要，经济性——从每种条件每个被试中都能获得一个结果	顺序效应 练习问题 被试猜测假设，并且按照假设行动 每种条件需要不同但相等的刺激（如词列表或数字问题）
配对	无顺序效应或被试猜测假设问题，不同条件的被试相等	完全匹配很少，因此某些被试差异可能使结果偏差 一个被试丢失意味着整对数据丢失

练　习

1. 为什么实验程序需要严格地标准化？
2. 为什么研究者想重复一个研究？
3. 独立样本设计中的被试在因变量上提供多少个分数？
4. 记得前面提到的草生长练习吗？我们来看两种评价生长的方法。这两种方法对应于哪种实验设计？
5. 重复测量的缺点是什么？
6. 配对设计的每个被试如何被分配到不同的实验条件？
7. 你的老师给你一组需要解决的字谜，一半是具体词，另一半是抽象词。记录解决每个字谜的时间。在这个教室实验中，什么是自变量？什么是因变量？使用了何种实验设计？特别需要何种预防措施？

8. 指导学生配对工作,研究是否我们用左手学习迷津快于右手。一个学生首先用左手,然后用右手学习。另一个学生先用右手然后用左手学习。请问这是什么设计?特别需要何种预防措施?自变量和因变量分别是什么?
9. 给予一组被试咖啡因,另一组被试纯净水,测量反应时,请问实验使用了何种设计?主要缺点是什么?
10. 在同一地点,分别在晴天和雨天进行卖艺表演。记录喂猴子的人数,预测是否在晴天会有更多人喂猴子。请问这是什么设计?自变量和因变量分别是什么?

答 案

1. 为了控制可能影响自变量和因变量之间因果关系的所有变量。
2. 为了检查效果的有效性或排除可能的干扰变量。
3. 只有一个,因为他们只参与一种条件。在重复测量设计每个被试在研究的所有条件上都提供一个分数。
4. 分两次测量草丛同一叶片,是重复测量设计。在两个不同地点,测量两个独立的叶片样本,是独立样本设计。"组内"和"组间"也有意义。请注意这些设计实际上并不是真正意义的实验,因为没有一个人来操纵自变量(如晴天和雨天)——下一章将会作出清晰的解释。
5. 顺序效应——常常通过平衡来处理。
6. 随机:每对中的每个被试被分配到任何一种实验条件的机会是均等的。
7. 自变量:词的类型。因变量:解决时间。设计:重复测量设计,并随机分配刺激。
8. 重复测量设计;平衡;自变量:手(左、右);因变量:解决时间。
9. 独立样本设计;在被试变量上可能不等组。
10. 独立样本;自变量:天气类型;因变量:人数。

关键术语

组间(between groups)
干扰变量(confounding variable)
平衡(counterbalancing)
独立组/样本/测量(independant groups/samples/measures)
配对(natched pairs)
非等组(non-equivalent groups)
顺序效应(order effect)
被试变量(participant variable)
前测(pre-testing)
相关设计(related design)
重复测量(repeated measures)
重复(replication)
标准化指令/程序(standardized instructions/procedures)
独立设计(unrelated design)
组内(within groups)

3.4 实验室实验和现场实验

在心理学领域谈起**实验室**,人们马上就会想到受害人被电线连接到精巧的装置上,以测量他们的想法。实际上,在心理科学里,实验室指人们被测试的地方,典型的是大学心理学系中的一个房间。一些心理学系会使用相当先进的设备,而其他的则可能进行面谈或纸笔测验。

相对而言,心理学家常常进行**现场**工作。这意味着他们走出研究室,在外部世

界里进行自己的研究，如学校、医院、办公室，甚至在大街上。这使研究者能根据人们的自然习惯来研究他们。这些人们可能知道，也可能不知道他们正在参与研究。他们可能被要求在医生手术室完成问卷，在工作场所被采访，作为街上实验者导演的抢劫案中的目击者而被询问看到了什么。

实验室实验的一个特征是，一旦进了实验室，被试就离开了自己的势力范围，并且明白他们正在被研究。在现场实验里，人们也知道他们正在被研究，但他们不必专门去到一个特殊而不熟悉的地方来参加测试（事情总是有例外的）。

> 我的同事卡拉·费安娜格喜欢用一些极小的问题来逗乐学生。设想一个被试按要求来实验室参加实验。当被试坐在等待室里等待时，一个富有魅力的男子坐下来寒暄了几句，然后明目张胆地从桌子里偷走10美元。实验目的就是看被试是否会报告偷窃。这里的主要问题是：这是现场研究还是实验室实验？
>
> 这个小问题正是关键所在。就被试而言，他们还处在正常生活中，尽管他们参加实验的时间是一个奇怪的日子。但是，就他们所知，实验还没有开始，因为心理学家还没有出现。这个情境与医生在手术室等待或进行抵押面谈相同——少见但属于真实生活。
>
> 不应当在这样的灰色区域来要求你在评估中回答问题。我提出这个问题，只是想集中于被试的心理状态。作为一个关键因素，被试心理状态决定了实验是否在实验室进行，并且强调一旦一个真正的实验室实验开始，无关因素就会介入。当然，我们知道被试在现场研究中也可能会意识到他们在参加实验。这种意识不仅仅是发生在实验室内，而且也很难理解被试会没意识到自己正处于实验室。

实验室实验的优点

- ❑ 在实验里，理想状况是控制所有的无关变量。因为一些心理学测试设备无法运到实验室外工作，所以当需要记录人们的高精确表现（如记忆、警惕性）时，在心理学实验室里是最容易进行的。
- ❑ 在实验室里，能仔细操纵自变量和仔细测量因变量。班杜拉（Bandura，1965）给精心挑选和指导良好的儿童播放影片，片中成人因攻击行为受到惩罚或受到奖励。之后，认真观察每个儿童在游戏室和与影片中相像的鲍勃玩偶在一起的行为。
- ❑ 在控制较少的环境中测量儿童的攻击性，观察者可以记录儿童在学校操场的活动。问题在于无法限制儿童：他们可以分开、聚在一起或隐匿在别人身后等。而在实验室里，可以尽可能地控制所有这些变量。

对实验室实验的批评

偏爱定性方法（见第7章）的心理学家批评说实验室研究只是从正常生活背景来抽取非自然行为的片段来研究。然而，几百位偏爱定量研究的心理学家反对在心理学实验室中研究的人为性。奈瑟尔（Neisser，1978）评论到那时为止的记忆研究：

我们已建立了稳固的以实验或观察为依据的概论，但大多数内容如此明显以至于10岁的儿童都至少知道……如果X是关于记忆的重要性或在社会上有显著作用的方面，那么心理学家几乎从不研究X。(pp.4-5)

- **测量的狭窄性**：问题在于，测量的需要和变量的精确控制导致对日常概念的测量过分狭窄。例如，班杜拉(Bandura,1965)的研究使用殴打玩偶作为因变量，但这种行为只是儿童能够产生的所有攻击性破坏行为中的一个很狭窄的样本。尽管我们从研究中知道，儿童将准确地复制成人的行动，但这几乎不是新闻，而且声称攻击是在日常生活中从他人习得的也几乎没有价值。与此相似的还有前面描述的鲁尔(Rule,1987)等人的攻击测量——为故事编写结尾，也把攻击的测量狭窄化了。
- **缺乏推广性**：一些效应似乎只是产生该效应的实验设计的结果。社会懒惰(social loafing)曾经被认为是一个普遍的心理学特征——人们处于一个团体时比单独时更不努力工作。但是，在作为一个一般的心理学现象被接受之后，跨文化争议出现了。在一些文化中，人们在团体中比单独时**更努力**工作(例如，Early,1989)。霍尔特(Holt,1987)的研究表明，甚至在西方实验室里，如果在要求人们进行团队工作之前只给予一点时间自我介绍，并没有见到社会懒惰效应。最初的实验是高度人为性的，当人们必须进行团队合作之前不允许有任何的互动。研究缺乏**生态效度**(ecological validity)——无法推广到其他情境，特别是从实验室推广到现实生活。使任务具有现实性常常是提高生态效度的一种方法——但这种方法并不总能奏效。
- **人为性**：对被试来说，实验室是一个非常令人害怕的地方，如果实验者仅仅给予正式的指导语，而没有通常人类社会互动的微笑和姿势，就更令人害怕。但是，对人为性批评常常聚焦于要求被试执行的任务的类型。例如，在早期社会懒惰研究中，要求被试拉绳子或制造噪声——相当狭窄和罕见的工作例子，但心理学家试图将结果推广到职业世界。当批评人为性时，要注意的是实验室环境**当然**是人为的。反对人为性的人却遗漏这点。在一般的科学里，我们粗略地观察现实生活中的事物，然后把这些事物带进实验室进行严格的测试，接着再把结果应用到实验室之外的现实生活中(见图3.4)。在医学研究中很普遍，但考虑经典物理演示，所有物体都以相同的加速度下落，但在现实中我们不可能看到——一片羽毛和煤以相同的加速度落下似乎是荒谬的。但是，一旦进入实验室，创设真空，让羽毛和煤片同时落下，你就会看到统一的加速度假设确实是真的。

实验室实验的优点和缺点概括如表3.3所示。

表3.3 实验室实验的优点和缺点

优 点	缺 点
变量可以被良好地控制和精确地测量	测量的狭窄性
能使用复杂的设备	一些效应只出现在实验室里，无法推广
排除了实验中被试行为的不可预测性	任务和人类行为测量的人为性

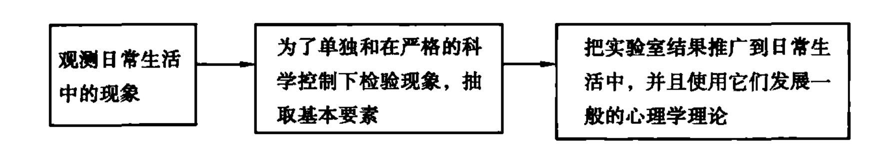


图 3.4　人为创设实验室的原因

3.5　实验效度——内部效度和外部效度

通过这章我们已经看到如果实验中的一些因素没有得到良好的控制，就可能导致从结果中得到不正确的结论。库克和坎贝尔(Cook & Campbell,1979)把这些因素称为**效度威胁**(threats to validity)。**内部效度**(internal validity)指诸如干扰变量、统计误差、数据收集偏差等因素。当效应原本不存在时，实验证明它存在，或者得出结论自变量产生了差异而这些差异原本由另一变量引起时，这些因素会给人留下印象。**外部效度**(external validity)指如果结果是真正的效应，它能被推广到其他情境(生态效度)、人群(总体效度)和时间(历史效度)的程度。图 3.5 列出了每类效度的相关问题。

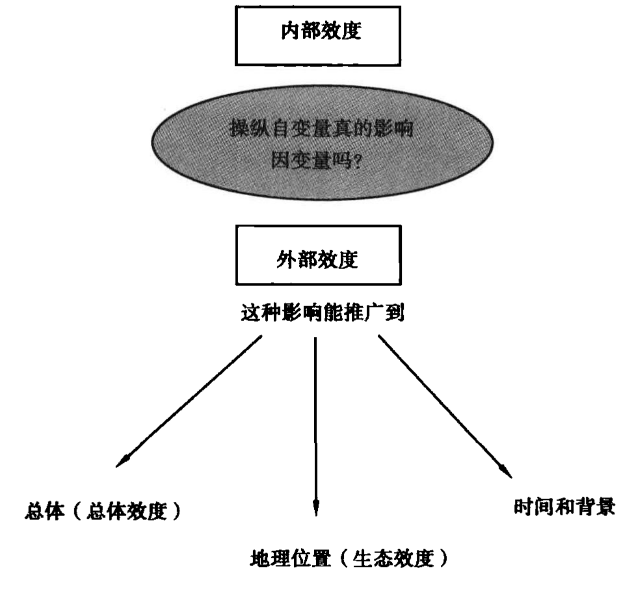


图 3.5　内部和外部效度

生态效度

因为生态效度被广泛使用，所以这个概念值得特别提及，有时在一些教科书以及当今教学中的使用是不正确的。生态效度指一个研究效应在多大程度上可以从一个地方或情境推广到其他地方或情境。典型地，采用流行的观点，认为实验室实验的生态效度比较低，因为无法在现实世界里执行它们。实际上，对现实性问题有一个很好的术语。卡尔史密斯、埃斯沃斯和艾诺森(Carlsmith, Ellsworth and Aronson,1976)使用**世俗现实性**(mundane realism)这一术语来表示实验室研究在多大程

度上模仿了现实生活中发现的事件和变量。他们使用**实验现实性**(experimental realism)来描述实验研究,实验情境是如此富有吸引力和魅力,以至于任何人为性都可由被试认真对待情境而获得补偿,尽管实验室情境与现实生活有很大的不同。米尔格雷姆(Milgram,1974)著名的服从研究就是一个例子。

近来的趋势指出,所有在自然环境开展的研究都比实验室研究有更高的生态效度。但事实并非如此,由于设计的草率性,许多"自然"研究产生的结果根本无法推广。另一方面,许多实验室效应确实可以从实验室推广到其他情境,米尔格雷姆效应就是极好的例子,这一效应可以在许多不同的情境中重复出现。另一个在现实的医院情境中有关护士的服从研究却从未获得重复,哪怕是局部重复也未产生服从效应(Rank and Jacobson,1977)。因此可以说霍飞林等(Hofling et al.,1966)人的"自然"研究生态效度低,而米尔格雷姆的研究生态效度高。

凯瓦拉希维丽和爱丽丝(Kvavilashvilli and Ellis,2004)认为一个实验可能脱离现实,但如果它的效应能够推广,就拥有生态效度。他们以艾宾浩斯记忆实验中的无意义音节为例。艾宾浩斯实验的材料和任务与现实生活中的记忆任务非常不一致,尽管其中混淆了许多其他因素,但他展现的效应能够在日常生活中发挥作用。在医学和生物学研究中也有相同情况,我们观测一个现象,进行高度人为的实验室实验(例如在培养皿中进行微生物的培养),然后在日常生活中重新解释结果,从而拓展了我们关于疾病的总体知识,并产生新的治疗方法(见图3.5)。但是在心理学中,通过让任务和情境更现实,我们有可能提高生态效度。然而,生态效度必须总是使用研究结果来评估,以表明一个效应会跨情境推广。我们不可能仅仅因为它的"自然性"来猜测一个研究生态效度是高的。

练 习

1. 在没有重读本章的情况下,思考使用实验室进行心理学研究的两个优点和缺点。
2. 下面每句话指的是什么效度?
   (1)如今不能发现阿希效应。
   (2)在许多非西方人群中没有作用。
   (3)在医院有效但在健康中心无效。
   (4)我认为你获得这些结果,因为在听众中有很多的戏剧专业学生;而且你的测量范围太狭窄。

答 案

1. 见表3.3。
2. (1)历史效度。(2)总体效度。(3)生态效度。(4)内部效度威胁的两个例子。

关键术语

生态效度(ecological validity)	内部效度(internal validity)
实验现实性(experimental realism)	实验室研究(laboratory study)
外部效度(external validity)	世俗现实性(mundane realism)
现场研究(field study)	总体效度(population validity)
历史效度(historical validity)	效度威胁(threats to validity)

3.6 实验和其他研究中可能存在的偏差

需求特征——反应性问题

被试在实验里具有**反应性**(reactive)。不像物理学里测试金属和石头,人类被试会对研究作出反应。实验情境是人类互动的一种,即使过去许多心理学研究所使用的古老语言极力想证伪这样的观点。实验对象(participant)被称为被试(subject),他们就像"跑动"的老鼠,被操纵在实验里。

但是,好奇是人类的本性,人类在诸如实验的陌生情境中,特别有可能寻找线索,从而弄清发生了什么。奥恩(Orne,1962)把这些线索称为实验情境的**需求特征**(demand characteristics)。这些线索帮助被试猜测研究者调查什么和期望什么。他们可能揭露实验假设。卡雷米斯等人(Carlamith et al. ,1976)认为由于要求被试全神贯注于实验程序,实验现实性可以减少需求特征的所有影响。

被试反应

如果需求特征的确起了作用,那么也存在**被试期望**(participant expectancy)的可能性。被试期望是指当被试期望某种特定的结果发生时所采取的各种行为。被试可能仅仅是为了让实验者感到高兴,从而制造出他们认为实验者所期待的结果。更隐秘的是,他们想表现得像其他人一样正常,例如涉及个人习惯的问题时,这称为**社会期望效应**(the effect of social desirability)。同样地,被试还可能受到**评价顾虑**(evaluation apprehension)的影响,即担心做得不好或表现得不符合社会期望,这种焦虑会影响他们在实验里的表现。

霍桑效应

20世纪20年代,在美国伊利诺斯州的西塞罗,西部电气公司的霍桑工厂开展了一项大规模的应用心理学项目。罗思丽丝贝杰和迪克森(Roethlisberger and Dickson,1939)报告了这个工作心理学的早期研究系列结果。研究者操纵了几个变量,包括照明、激励、休息间隔、工作日长度。针对后面两个变量,有5名工人在单独房间里被观测。研究发现大部分情况下生产率在条件变化后提高了,甚至在条件返回到最初条件时生产力仍然提高。这样的结果可能是因为置于单调的车间条件的工人,由于转到新的社交性情境或独立环境中,从而导致了产量的提高。当研究者通过实验改变了薪水和奖励系统时,效应也受到了混淆。然而,研究方法中仍然采用了**霍桑效应**(Hawthorne effect)这一术语,用来描述这样的情境:当被试知道他们正处于实验观测中时,他们的行为就会受到影响。

还有一个因素与需求特征有关:**启发**(enlightenment)。心理学学生越来越清楚心理学的研究结果,以及研究者所使用的策略,即通过伪装研究意图来获得真正的结果。由于70%的研究被试是学生,很多还是心理学专业的学生,因此启发便是一个重要的因素了。

实验者期望

> 罗森塔尔(Rosenthal,1966)在研究中给学生分配老鼠样本,并告诉一些学生他们的老鼠是"笨"的,告诉另一些学生他们的老鼠是"聪明"的。令人惊奇的是,老鼠在迷津学习中的表现和给予它们的标签一致:聪明老鼠学得更快。我用"惊奇"一词是因为老鼠实际上是被随机分配到两组。所谓"聪明的"根本不是特别聪明。

你认为在这个研究里发生了什么?这一切之所以发生,是因为学生期望聪明老鼠有聪明的行为,但这并没有真正告诉我们这一切是如何发生的。**实验者期望**(experimenter expectancy)指实验者因知道所期望出现的研究结果,而潜移默化地改变被试行为的可能性。但是,期望纯粹是一种心理状态。我们寻找的是研究者可能影响老鼠表现的某些实际行为。巴伯(Barber,1976)认为学生们也许从他们遵循的实验程序中获得了实验意图——但这是为什么呢?

我们必须承认,在充满压力的研究世界中,年轻、充满激情,但也许有点脆弱的实验者试图以他们喜好的方式来接近实验效应。(顺便说一下,**实验者**指的是代表研究者来运作实验的人,或者指的是负责整个研究项目的**调查者**)。然而,对于实验者期望的研究并没有集中于**有意**偏袒结果的问题,而且更多地集中于这样的意见:对于实验期望结果的了解会使实验者的行为泄露实验的秘密,或在某种方式上影响他们的行为。

罗森塔尔和雅克布森(Rosenthal and Jacobson,1968)的著名研究表明,只要让教师相信一些学生经过一个教学年度发展,智力会有显著增加,这些学生便真的会有这样的发展,即使学生的名字只是随机挑选的。但是,从1968年到1976年,在对这个现象的兴趣最为浓烈的时候,有40项实验并没有展示出实验者能产生与他们的调查者所期望的方向相一致的结果。然而,文献里也存在不寻常的事情,正如下面的这个例子:

> 威格、斯托特和克奥拓斯(Wigal, Stout and Kotses,1997)要求新手实验者收集数据,被试被描述成可能会也可能不会对呼吸困难的暗示作出反应。测量被试的呼吸阻力(呼吸限制的技术测量)。可以肯定,被描述成有可能反应的被试比另一组表现出更高的呼吸阻力,但对于"可能会"和"可能不会"的描述是随机的。

为了预防因实验者期望而产生的不必要变异,以及检查实验者一般的数据收集能力,研究者会使用**实验者信度**(experimenter reliability)这一测量指标,即统计分析实验者结果的一致程度。

调查者效应

一些人使用术语**调查者效应**(investigation effect)来指调查者对研究结果产生的所有的不必要的影响,包括期望效应、给被试的无意识提示等,也包括**研究设计中**采取的行动,如被试选择、材料、条件顺序、指导语、刺激等。因此,调查者效应是一个上位概念,也许在单独分析引起研究偏差的所有分离的效应时更有用。

处理期望问题——双盲和单盲

如果我们不让人们知道期望的结果是什么,似乎我们就能避免期望效应。如果不告诉被试他们所处的实验条件,那么这种情况就是**单盲实验**(single blind)。例如,不告诉被试他们服用的是咖啡因还是安慰剂(如盐,对行为根本没有影响)。当实验者和被试都不知道实验的期望结果时,就称为**双盲实验**(double blind)。

3.7 被试意识——不只是实验室实验才有

一些人认为主要是实验导致了所有问题:反应性、期望、需求特征等。但是,想一想你会更加清楚明白,在任何研究里,只要被试意识到他们正在被研究或他们可能清楚期望的结果,就会发生这些相同的效应。在操场被观测的儿童会尽力表现得高兴;在检验工作条件的改变是否会提高生产率的研究中,被观测的工人可能会按照理论的期望去工作。如果被试知道他们正在被研究,并且能够猜测研究者的预期,那么所有的这些因素都会起作用,不管是否是实验研究,也不管研究是在哪里进行。

练　习

1. 复习一下40-41页中的练习7-10,哪一个研究里有需求特征的影响?
2. 哪一个研究能使用单盲程序?
3. 为什么不是所有的实验者都获得同样的结果?

答　案

1. 需求特征都有影响,除了练习10的被试不知道有实验者。
2. 所有。单盲程序与练习10的被试无关,因为他们不清楚实验。但是人数记录可由"盲"观察者进行。
3. 有非常多的原因。一些人可能泄露给被试更多线索;一些人无意间让结果倾向于假设方向;一些人迫于现实压力而尽力发现(但我认为这种现象很少);一些人可能更令被试害怕;一些人草率地收集数据或统计分析;还有一些人令被试更渴望高兴——我也想知道原因呢!

关键术语

需求特征(demand characteristics)	调查者(investigator)
双盲实验(double blind)	调查者效应(investigator effect)
启发(enlightenment)	被试期望(participant expectancy)
评价顾虑(evaluation apprehension)	安慰剂(placebo)
实验者(experimenter)	取悦实验者(pleasing the experimenter)
实验者期望(experimenter reliability)	反应性(reactivity)
实验者信度(experimenter reliability)	单盲(single blind)
霍桑效应(Hawthorme effect)	社会期望(social desirability)

4

非实验研究方法

本章内容

- ❑ 探究准实验的含义。这类结构化的研究多数在现场进行,因为研究者缺乏对一定条件的控制,所以并非真正的实验。
- ❑ 介绍各类非实验研究,包括观察、访谈、调查以及个案研究,在随后章节中将对之作进一步解释。
- ❑ 探讨横断、纵向和跨文化研究的设计。

4.1 准实验

上一章我们提到,若选择过多攻击型被试到炎热的房间,那么鲁伦等人(Rule et al., 1987)有关攻击性与温度关系的研究结论就会被曲解。实验组攻击性更强的结论,则可能是由于该组被试整体上本来就更好斗。而在此实验处理下,高度对结果的影响或许是微乎其微的。这就是非等组的问题。通过将被试随机分配于各种实验条件,实验者尽力使非等组导致的问题最小化。然而,有时候这根本不可能实现。

现场实验通常限定某组为实验组,另一组则为控制组。例如,某心理学家在一工厂做研究,考察生产线速度对工作表现的影响。这时无法将工人们随机分配到各类实验处理中,因为不可能打乱现有的轮班工作组。最好也是最可能的办法,就是分别考察低速的某轮班工作组及高速的另一轮班工作组的表现(见图4.1)。但问题是,不同实验处理中的压力水平差异可能在此之前便存在——低速轮班工作组的工人可能实验前本来就处于更低的压力水平。

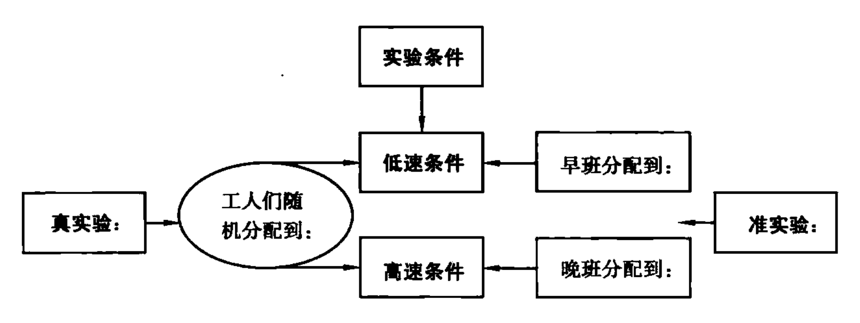


图4.1 真实验和准试验设计

正如我们之前所见,在真实验中研究者操纵自变量,并使其他变量恒定,包括随机分配被试到各种处理中。在准实验中,也有可辨别的自变量,因变量的变化可以通过严格操作的方式观察到。然而,该类实验并不符合真实验的严格标准(因此被标为"准"实验)。准实验研究有两种典型特征:

1. **被试并非随机分配** 有时,某类实验设计根本不可能随机分配被试于各种处理,尤其是现场实验。加尔丁、罗农、卡尔格林(Cialdin, Rono and Kallgren, 1990)在路上随手丢下大量垃圾,然后观察接到传单的不知情的被试是否会将传单扔到地上。当地上的垃圾越来越多,而被丢弃的传单也越来越多,则表明人们在随大流。在这种情况下,地上出现更多的垃圾似乎在暗示他人,扔传单是可以的。在该研究中,研究者并没有控制被试在自变量上的不同水平——地面上更多或更少垃圾。
2. **研究者对自变量没有加以控制** 实验并不总是研究者设计出来的,有时生活中发生的事情也可以用来做实验。例如,某教育当局可能决定选择一些学校进行反欺负行为教育。研究者没有对实验加以控制,因此变量没得到心理学研究者所希望的控制。尽管如此,心理学家仍可以利用此种情形来调查该教育对儿童欺负行为的影响。根据所属区域、平均经济水平和父母受教育程度等特征,研究者可以找一所没开展反欺负行为教育的平行学校,这样对这两所学校可以尽可能公平地进行比较。

自然实验

准实验还包括**自然实验**(nature experiment)——生活中常出现这样一些情形,没有安排特定的实验,更没有上述例子中那样的专家控制,但可进行各种情形的比较。罗斯、坎贝尔和格拉斯(Ross,Campbell, and Glass,1973)研究了1967年英国引进呼吸分析器前后司机们的行为变化。经过对交通事故的统计分析,他们认为超速驾驶现象越来越少了。通过分析酒类销售、事故率以及事故次数等数据,排除了其他假设。[①] 沃德和沃特斯在(Wardle and Watters,2004)最近一项有趣的研究表明,在校园里与年长女生接触增强了年幼女生对体重和饮食的消极态度。(见表4.1)

表4.1 女孩类型和沃德与沃特斯(Wardle and Watters,20004)的饮食态度研究结果

类 型	年 龄	学校类型	年龄最大的学生
接触型女孩(exposed girls)	9岁	初中	13岁
	11岁	中学	18岁
非接触型女孩(non-exposed girls)	9岁	小学	11岁
	11岁	初中	13岁

总体上,"接触型"女孩有更苗条的标准,感觉自己更超重,结交了更多有过节食经历的朋友,在儿童饮食态度测试中得分更高,自尊心更低。

这是一个很巧妙的设计。因为"接触型"和"非接触型"女孩样本年龄相同,而实验处理不同。"接触型"女孩在学校以比自己大4岁到7岁的同辈为榜样,而"非接触型"女孩的榜样最多比他们大2岁。

4.2 准实验的麻烦

对准实验的关注点是缺乏控制所造成的"效度威胁"。控制越少,对观察结果的可能解释就越多。库克和坎贝尔(Cook and Campbell,1979)开展了效度与准实验的争论。那时他们和其他人所面临的问题是,与其他研究方法相比,心理学研究者们更信任和尊重真实验研究,而认为大多数其他研究方法得出的结果更糟糕。与其他在诸如教育、健康或体育等应用领域工作的心理学家一样,库克和坎贝尔也发现,在现有群体和现有条件下很难进行真实验,很难进行控制良好的研究。他们认为,如果其他因素控制得好,对数据分析再作些调整,那么此类实验应该得到认同,因为它们同样可以为心理学知识作出有价值的贡献。有关问题的详细论述可参见库里坎的研究(Coolican,2004)。

4.3 非实验研究的常见类型

大量心理学研究,尤其是那些远离更科学、可控领域(诸如生理心理学或认知心理学)的研究,既非真实验,也非准实验。对所谓的准实验应该有个界限。一般来说,研究者能(在相对较短时间内)判断出各组人群处于自变量的不同水平之下,

---

① 是呼吸分析器的使用而不是酒精销售的变化,导致了超速驾驶下降。——译者注

而这些水平通常被看做“处理”。据此,我们不能把一项考察内外向型性格者或高低焦虑者之间的差异研究称为准实验。在此情形中,并没有任何简单的“处理”。有些人将性别差异研究称为准实验,但实际上这远远偏离了准实验概念。将男性或女性看为“自变量”的一个水平不能解释为实验意义上的“处理”,因为这涉及太多未加控制而又令人困惑的变量——成千上万!男女在大多数变量上得分几乎一致,包括在智力测验上。近来在英国,数学和科学方面原本认为处于劣势的女生,已经赶上了男生,而且普遍表现得更好。而要找出诸如逆向停车方面的性别差异,其本身通常就不太有意义。“性”或性别因素实在是太宽泛了。

专业术语综合征

在第2章中我提到学生(以及导师)经常表现出“专业术语综合征”,即认为有必要让事物都有一个名称,如果同一概念以另一个名称蹦出时,就会让人感到不自在。我认为为研究设计找一个“正确”的名称就是这种表现。人们想方设法为各种研究设计找到相应的名称。事实上很多设计不止一种名称,而考试委员会则使我们尽力将各个研究分门别类,并确定其名称。除实验外,你也会发现,在各式各样的不同水平的教材中,同一研究有多种称呼。不要担心!写报告时,你不必找到一个特定的专业术语,只要描述出有关变量及其主要程序就可以了。回答考卷问题时,只要了解老师的习惯以及考试委员会的偏好就行。尽管在非实验研究领域没有绝对正确的答案,此时我们仍将就一些术语和说明作些阐述,以便为接下来的章节作好铺垫。

事后研究

如果比较两个及多个预先存在的群组(例如内倾型和外倾型),那么我们可使用组差研究这一术语;或更一般意义而言,若所测事件和特征已存在,对于这样的情形,我们可以使用**事后研究**(post-facto studies)的术语。在诸如下段有关焦虑和自尊的例子中,不需要实验,只需测量人们已有的变量或已发生事件的影响就可以了。

相关研究

有些教材将所有非实验研究称为“相关的”,但问题是许多此类研究事实上并没有使用相关分析。真正的**相关研究**[correlational study,定义为“采用相关分析的调查”,见英国资格评估与认证联合会(AQA)的英语A级考试大纲]则试图考察焦虑是否与自尊有关。人们或许会假设自尊强的人不可能是焦虑者,而自尊弱的人会更焦虑(事实上可能是:焦虑会降低个体的自尊,而相关研究并没有告诉我们谁是因谁是果,见第12章)。我们不妨挑选被试,让他们完成两份心理量表,分别测量焦虑和自尊。我们既不是在操纵自变量,也绝不是在进行一个实验,我们只不过在测定一个既定事实,然后我们运用相关的统计程序来测量这两个变量之间的关系强度。

观察研究

同样,有些教科书称所有的非实验研究为“观察性的”,理由是研究者在情境中

并没有进行干预，而是用这种或那种方法进行观察。然而，一般说来，唯有当研究的主要方法是观察法时才可称为**观察研究**(observational studies)。我们将在第5章中详细讨论。

询问研究

询问研究(studies using questions)是一类研究的总称，是一种主要使用访谈、问卷和调查法的研究，也可能包括一些个案研究(对某个或几个被试提大量的问题)。除了非实验研究的共同缺陷外，其余问题点到为止，有关详细阐述将在第6章和第7章中进行。上述个案研究，以及相关研究和观察研究，都具有非实验研究的共同缺陷：无法确定因果。

非实验研究的缺陷——因果关系

在第1章，我们看到，当研究人员报告结果似乎支持某理论时，总有"啊，但是"这样的疑问。假如我们正进行一项研究：观察儿童观看暴力倾向电视节目的时间，或问他们诸如"你喜欢《辛普森一家》中的伊曲和斯科拉奇么?"①。无论哪种情况，我们都假设：观看暴力倾向电视会导致儿童更高水平的攻击性行为。现在我们观察儿童在操场或其他形式的自由玩耍的行为，然后对这两种测量结果进行相关分析。结果发现，观看暴力倾向电视更多的儿童在玩耍时肯定表现出更强的攻击性，反之亦然。我们呈现的这些调查结果支持以下结论：观看暴力倾向电视节目引发高水平的攻击性。

然而，假如两者的关系是另外一种情况呢？你怎么知道不是遗传使他们比其他儿童更喜欢看有暴力倾向的电视节目呢？这就是确定因果方向的问题(见图4.2)。

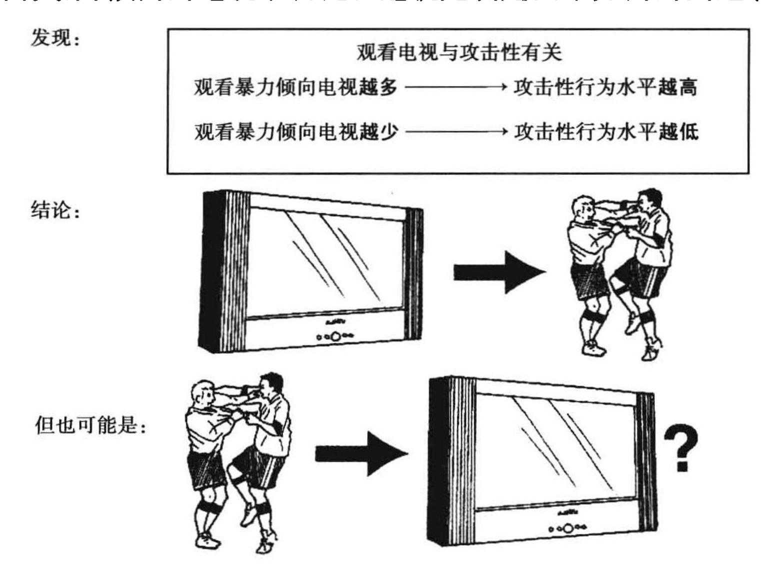


图4.2 非实验研究中的因果关系问题

① 《辛普森一家》是一部热播的美国喜剧动画片。伊曲和斯科拉奇分别是该动画片中的一只猫和老鼠，拍摄《辛普森一家》的福克斯公司用伊曲和斯科拉奇讽刺迪斯尼公司的《猫和老鼠》是主张暴力的动画片。——译者注

正如前面我们所看到的,实验几乎不受这种解释的影响,因为实验者**操纵**假定的原因,然后观察其影响。准实验在判断因果关系时稍微弱些,即任何所发现的影响,可能是被试间差异所致,也可能是研究者没有控制自变量的情境差异所致。

纵向研究

要解决因果方向问题,方法之一是进行**纵向研究**(longitudinal studies)。在这种设计中,对被试组进行长期研究——通常为时数年。在此,如果在儿童表现出任何有攻击迹象前,我们观察儿童观看暴力倾向电视的情况,然后发现数年后他们更具有攻击性,那么我们就更加相信,观看暴力倾向电视节目影响了随后的攻击行为。当然,也有可能是遗传因素[①]使儿童更多地选择观看暴力倾向电视,进而导致随后攻击行为的出现。然而,如果对两组进行研究,一组由父母控制暴力电视节目的观看,而另一组则没有,那么因果方向更有可能是由观看电视节目导致攻击行为,因为儿童攻击本性不可能导致父母控制观看电视的差异。

注意,如果用这种方法跟踪一个相当大的儿童群体时,该群体就称为**同期组群**(cohort)。目前,瑞典已有许多全国性的同代研究(cohort study),例如俄克尤斯、克里特森和赫杰恩(Ekeus,Christensson and Hjern,2004)跟踪研究了1987年到1993年出生的800 192名瑞典儿童,直至他们长到7岁。他们发现,除其他方面的原因,少女妈妈的孩子早期伤人的概率显著比其他儿童大,包括蓄意伤人。

电视真的会让你具有攻击性吗?

厄伦、休斯曼和莱夫克维兹(Eron,Huesmann and Lefkowitz,1972)对9岁的儿童进行了跟踪研究。正如其同龄儿童所判断的,这些儿童对收看暴力倾向电视节目的偏好与其攻击性相关。研究者利用其他一些统计技术发现,10年后这种关系甚至更强了。厄伦等人在收看暴力倾向电视与随后的攻击性行为之间建立起一种极有可能的因果联系。

一些纵向研究在准实验中就是以这种方式设置控制组的。卡根、克斯里和泽拉佐(Kagan,Kearsley and Zelazo,1985)的研究表明,与经常呆在家里的儿童相比,如果托儿所还不错的话,幼儿期大部分时间在托儿所度过的儿童在发展中没有什么障碍。

横断研究

纵向研究有一大缺点:研究者必须等待多年才能获得结果。想必会有更快捷的方式吧?假如我们对儿童心理理论的发展时间感兴趣(见表4.2)。心理理论指个体推测他人在特定情境下的意图的能力。

为了估计心理理论发展的平均年龄,研究者横向测试了从三岁起到更大的儿童(例如,温默和帕那(Wimmer and Perner,1985)以3~9岁的儿童为研究对象)。年龄是用得最普遍的横断面,但是**横断研究**(cross-sectional study)总体上是用来比较不同人群类别的测量结果,例如职业,教育类型和少数民族群体等。年龄分类的最主要优点是,可以很快获得纵向研究需要花很长时间才能得到的结果。然而,纵向

① 某些儿童先天具有攻击本性。——译者注

专栏 4.1 检验思维推理——错误信念任务(Perner, Leekam and Wimmer, 1987)

为了检验儿童是否发展起心理理论,研究人员可能会使用筒装聪明豆,并将之展示给儿童,问他们里面装的是什么。一点都不奇怪,儿童通常回答是"聪明豆"。接下来,研究人员给孩子们看,筒里其实装的是铅笔,并问他们一个关键的问题,"如果詹姆斯(一个朋友)走进房间,他会认为筒里装的是什么?"那些未发展起心理理论的儿童不能判断詹姆斯会看见什么,他们往往根据自己刚刚了解到的来回答这个问题,因此他们会说"铅笔",而发展起心理理论的儿童则更现实地回答"聪明豆"。

研究跟踪相同群体,而横断研究存在的问题是每次测量的是不同的群体,所以群组间可能是非等组。

跨文化研究

专栏 1.2 的例 5 很有趣。在此,我们建议通过观察墨西哥和冰岛司机来检验这种假设:炎热会增加攻击性行为。该研究似乎假定,原则上各地司机的驾驶行为都一样。如果在墨西哥发现了更多的攻击性行为,那就支持我们的假设。然而,这取决于我们用什么方式测量攻击性驾驶。我非常肯定,若把鸣笛次数作为测量方式,那么墨西哥司机显得更有攻击性。到诸如罗马和孟买旅游的北欧游客,常惊讶于城市司机们鸣笛不断的攻击性。但是如果认为鸣笛更多就意味着攻击性更强,这就显得很武断。在鸣笛频繁的国家,鸣笛往往不过是提醒别人你在那里。事实上,在英国这是你鸣笛的唯一合法理由,而在愤怒时非常频繁地鸣笛就会被认为是攻击性行为了。如果这在罗马或孟买被看做愤怒的信号,那么我们不得不得出这样的结论:驾驶者们一直都非常愤怒。显然,事情并非如此。在英国和印度,鸣笛的功能和目的各不相同,因此这种攻击性测量方式在这两国并不等同。所有社会鸣笛都是攻击性的观点,是西方种族中心主义的范例。该术语是指以自己群体的标准来判断其他群体的行为。过去,东方人常被认为是"卑躬屈膝的",因为他们打招呼时总是鞠躬,这其实是西方对东方传统习俗的错误诠释。

跨文化研究(cross-culture study),旨在通过不同文化之间的比较,考察在西方发现的效应是否在其他地方也会发生,或者说是研究心理学的"普适性"——在各种文化中都会发生的事情。我们也可以做些调查,来检测西方文化普遍存在而其他文化缺乏的变量假设。例如,面对两维的绘画而没呈现相应的三维描述,人们的知觉是否会受视错觉的影响?在西方有很多这类绘画,但在其他一些地方,人们则很少有这种经验。同样,在近来才普及电视的地方,人们的行为是否也会因观看电视而发生变化呢?[①]

很显然,进行诸如此类的比较时,若仅使用为西方人设计的测量手段,例如智力测验和人格测验等,将有可能产生很多偏见。有一个有名的故事,说波多黎各人在"一个男孩雨中撑伞角度不正确的 IQ 测试"题目上的得分为零。为什么?因为在波多黎各,男孩撑伞被视为太女人气了!

早期的跨文化研究方法受到了人类学家路斯·班底克得(Ruth Benedict)等人

① 在西方国家,儿童的行为深受电视节目的影响,但在电视刚普及的国家,情况不见得如此。——译者注

的启发。班底克得认为，只有通过观察个体所处的文化环境，才能真正理解他们的行为和思想——这就是大家都熟知的"文化相关"的观点。其他跨文化心理学家则采取更为普适性的立场，他们试图表明存在普遍性的心理学维度，而西方研究只看到维度的一端。我们可以考查其他文化是如何用完全不同的方式，往往以维度的另一端，展现出相同的心理构想。例如，史蒂贝克（Stipek，1998）的研究表明，美国学生偏爱对自己取得的成就感到自豪；相比之下，中国学生更倾向于对他人（例如朋友）取得的成就而非自己的成就感到自豪。对许多亚洲文化来说，张扬不合乎其习惯。

若想了解更多的跨文化心理学研究方法和结果，不妨阅读下列出版物：

1. Berry et al. (2002). *Cross-cultural Psychology: Research and Applications*(2nd edition)
2. Richards(1997). '*Race*', *Racism and Psychology*.
3. Shiraw and Levy(2004). *Cross-cultural Psychology: Critical Thinking and Contemporary Applications*.
4. Smith and Bond(1998). *Understanding Social Psychology Across Cultures: Living and Working in a Changing world*(2005) and the *Journal of Cross-cultural Psychology*, published by Sage Publications.

练 习

1. 下列哪些研究是准实验，哪些是自然实验，哪些根本就不是实验？同时指出哪些研究是现场研究。
   (a) 在实验室的反应时仪器上测试吸烟者和不吸烟者。
   (b) 某心理学家将儿童随机分配到幼儿园的两种早期阅读方案中，3 个月后记录他们阅读取得的进步。
   (c) 有两家非常类似的养老院，其中一家由国有（NHC）变成私有，评价这两个养老院的员工在变化前后的工作满意度。
   (d) 随机挑选一组曾经受轻微挫折的儿童，在实验室观察他们的攻击行为。
   (e) 将写有外国或当地人名字的钱包丢在电话亭里，看是否有更多的写有当地人名字的钱包被归还。
   (f) 在停车场观察男性和女性是倒车还是笔直开进泊位点。
2. 一项研究是准实验而非真实验的两大条件是什么？
3. 阐述一下横断研究和纵向研究的优缺点。

答 案

1. (a) 不是实验，是对现有群体差异的事后研究。
   (b) 现场实验。
   (c) 自然实验（因此为准实验）。
   (d) 实验室实验。
   (e) 准实验，研究人员没有控制被试的分配。
   (f) 不是实验，是对群体差异的事后研究。
2. 不能随机分配被试于各种处理中；研究者没有控制自变量。
3. 横断研究的优势：迅速获取不同年龄群体的结果；劣势：群体不等同。纵向研究优势：跨时间研究相同的群体，不存在非等组群体的问题；劣势：要花很长时间才能取得结果。

关键术语

同期组群(cohort)	相关研究(correlational study)
跨文化研究(cross-cultural study)	横断研究(cross-sectional study)
组差研究(group-difference study)	纵向研究(longitudinal study)
自然实验(natural experiment)	观察研究(obserational study)
事后研究(post-facto study)	准实验(quasi-experiment)

5

观察法

本章内容

- ❑ 我们不仅把观察法视作一种简单的技术，也把它作为一种主要的研究设计。
- ❑ 我们审视开展观察研究的场景，并考察这种场景如何影响人的行为，同时再来看一种特殊的设计——自然观察。
- ❑ 我们将看到结构化（或非结构化）观察，并且看到数据编码的方法。
- ❑ 我们还将谈到参与观察，在这种观察中，观察者本人可能成为被观察团体中的一员。

5.1 任何研究都是观察

所有的科学研究都必须利用观察法搜集数据。在大多数精确实验中，人们在进行测量和读取仪表数据时，观察是不可或缺的。那么，观察法对于心理学研究者又有什么特别之处呢？在实验室或其他严格的实验研究中，其潜在的问题是，个体所处的环境导致了他相应的行为。在一项任务中，研究者很少要求被试在日常生活中去记住一串无意义音节，相反，研究者可能对人们在一般情况下，比如超市购物或面试表达中所使用的记忆策略更感兴趣。观察使我们看到了日常情景下人们的行为。

5.2 作为技术的观察

我们可能并不想知道人们的日常行为，却很关注个体在特殊的情形中会采取什么行动。我们可能并不想测量个体的反应，却想知道他们做某件事情的次数。因此，在一个实验的限制下，我们可以把观察作为一种技术来使用。例如，班杜拉1965年的著名研究（Bandura，1965）就是一些观察实验，在这些研究中，他对儿童玩波波洋娃娃的行为进行了观察。这些研究是在一个实验室中进行的，其中一种可操纵的自变量是成人示范者的攻击行为是否受到奖赏。因变量则是儿童被观察到的模仿性攻击行为，并记录孩子模仿成人行为的次数。不过，这些记录是我们在控制情境中通过观察获得的。

5.3 作为研究设计的观察

如果一个研究的**设计**是以观察法为主导的，那么，它就是相对于控制更加严格的实验设计而言的。当他们在日常情境中使用观察法时，它所强调的是对被观察个体或团体的相对自然出现的、未经控制的行为的观察，在观察过程中，可能用到他们的知识，也可能不涉及他们的知识。

5.4 观察研究的种类

观察研究至少可以从三个主要维度进行分类。

1. 依据被观察者所处的情境是自然的还是结构化的。
2. 依据所搜集到的观察数据的结构化程度。
3. 依据观察者融入被观察者的程度。

让我们依次对这些内容进行讨论。

观察的情境

在班杜拉1965年（Bandura，1965）的实验中，观察的对象是在一个游戏室中玩一定数量物品的孩子们，孩子们的反应被严格编码，见下面的解释。这也正是为什

么班杜拉的研究通常被称作**实验**的原因。在这项研究中，有可操纵的自变量和通过观察（作为一种技术）来测量的因变量。因此，所观察的行为是发生在十分严格的结构化的实验情境中。相比较而言，在津巴多（Zimbardo，1972）著名的监狱情境模拟研究里，"监狱"中的事件是按常规发展的，而卫兵和囚犯的非限制行为则被全部记录。在这项研究中，被观察者可以做什么，在事前几乎没有什么限定。事实上，正如大家所知道的，由于对行为的控制过于宽松，以致对最终发生的事件失去了控制，模拟练习也不得不停止。由于本研究中卫兵的攻击性太强，囚犯具有受到严重心理伤害的危险。然而应当承认的是，津巴多的研究完全是在心理"实验室"的情境中进行的。

如果研究情境中所记录的是人们自然发生的行为，这样的研究一般被称做**自然观察**（naturalistic-observation）研究。在这样的研究中，被观察者处于他们日常生活情境中。一个典型的例子就是，训练有素的观察者在学校操场上对男孩和女孩所表现的攻击行为进行计数。在这里，孩子们将像以往一样在玩耍时表现出正常的行为，而观察者则运用某种技术以客观的方式对所发生的行为尽可能精确地加以记录。

自然观察研究的主要优点是，研究者可以观察到自然发生的、非限制性的行为，这些行为不会由于研究者的某种需要而受到干扰与扭曲。在班杜拉的研究中，行为受到实验情境和孩子得到的有限玩具的限制。在实验室研究中，人们所做出的行为是由于研究者要求他们那样做。而在自然观察研究中，我们所观察的是人们自然发生的行为，这些行为与个体知道自己被观察时所表现出来的行为是有明显区别的。我们接下来将探讨**非隐蔽观察**（disclosed observation）的一些问题。

练　习

1. 请说出研究者设计研究时使用观察法的几个理由。
2. 通过和"实验室"中的观察作比较，请列出自然观察的一些优点与不足。
3. 请描述班杜拉关于社会行为模型的几种假设是如何通过自然观察来观测的。

答　案

1. (a) 研究者想观察和记录非限制的自然行为，而不是那些由实验员的指导语所规定的行为。
   (b) 研究者想记录真实的行为，而不是人们在访谈中所报告的内容，因为在这样的访谈情境中，**社会期望**（social desirability）可能导致与实际情况不符合的场面出现。
   (c) 观察法可能是唯一的现实的选择。譬如，如果课题的研究兴趣是婴儿对养育者的依恋行为，或者是他们对语言的本能运用，那么直接观察似乎就是最有用和有效的方法。
   (d) 考虑到伦理原则会限制实验法的运用，因而可以把观察法作为最好的选择。例如，我们不能直接制造残疾，但我们可以观察刚刚残疾的人们是如何适应新的躯体的及其以后的进展如何。
2. **优点**：被观察者的自然行为不会因为焦虑情绪和为留下好印象而受到干扰，尽管这取决于研究者如何进入观察情境的；自然观察包括了行为的全部情境；当实验法受到伦理原则的限制或者当个体无法合作（譬如，一个伤病房中的父母，尽管这时候也允许这样做）时，自然观察法就显出了它的优势。

缺点:对无关变量缺少控制;对观察者的训练可能花时很长,代价很高;无法使用不便携带的装备;观察者很难一直处于谨慎和隐蔽的状态;如果一个编码系统已经使用而且难以改变,观察者就可能无法记录有意义的相关行为。

3.答案是开放的。提示:可以记录孩子在家里看电视时模仿行为何时发生。

关键术语

自然观察(naturalistic observation)

观察技术(observational technique)

观察设计(observational design)

观察的结构

结构化的观察设计有时也称为**系统观察**(systematic observation)。其目的是对行为作尽可能精确和一致的记录与分类。研究者在使用观察法时记录数据的方式有几种。使用观察法但同时使用某种**质性**方法(在本书第7章讲到)的研究者主张,对行为的简单分类可能会失去社会场景中行为的丰富性及其隐含的意义。我们将在第7章详细地讨论这一问题。在这里,本章主要集中讨论观察研究中传统的定量方式。不过需要指出的是,**参与观察**(participant observation)(在下面将讲到)常常从定性的角度去使用。在这种情况下,强调的不是对行为的分类与计数,而是对被观察到的行为所代表的意义的分析。

观察者常常用像图5.1这样的检查表来记录即时生成(或现场情境)的数据。他们可以使用一种单向玻璃镜,这样被观察者就不会意识到自己在被观察,也不会做出假如他们知道自己被观察时的相应行为(见下面)。当然,这对研究者来说就涉及伦理上的问题了。通常情况下,在观察完成之后,研究者要征询被观察者的意见,即他们是否同意把观察资料作为真实的数据保留下来,或者是希望销毁这些数据。研究者也可以使用某种录像方式来记录行为,不过像前面一样,在其他任何人看到这些资料之前,研究者同样要就资料的去留征询被观察者的意见。有时候,被观察者可能会知道自己被观察,但是单向玻璃镜的运用使他们不至于对自己作为观察目标感受过于强烈,从而也能更放松一些。

如果要进行统计分析,那么数据的搜集无论如何都必须使用**编码系统**(coding system)。如果观察者曾经接受过使用编码纸的训练,这些工作也可以在现场完成。图5.1是在操场攻击行为观察研究中使用编码纸的一个例子。

儿童编号	无缘无故地打或者猛推	学伙伴打或者猛推	报复性地打或者猛推	无缘无故地向同学喊叫	学伙伴向同学喊叫	报复性地向同学喊叫
A						
B						
C						
⋮						

图5.1 对攻击行为进行观察的编码系统

在这个编码系统中,事件发生的频次将被记录,即对被观察儿童喊叫的次数进行记录,不管他的喊叫是无缘无故的,还是报复性的,如此等等。在有些编码系统

中，每一个具体行为所持续的**时间**也可能被记录。观察者也可以用一种量表来测量某些指标，例如，一个被访谈者所表现出来的紧张程度。

不管观察者对现场发生的行为进行编码，还是通过观看录像记录对行为进行编码，遇到的共同问题就是**何时**观察行为。对一小时的录像进行编码可能会花费数小时，这就涉及取样的问题，因为，在操场上的观察者不可能在所有时间里观察到每一个孩子。对于取样标准，有几种可供选择的方案。**时间取样**（time sampling）指的是在一定的时间内观察一个特定的儿童或团体，例如，每5分钟观察15秒。**时点取样**（point sampling）指的是在某个时段的某一点上（例如，在每30秒的结束时）对每一个个体进行观察。**事件取样**（event sampling ）聚焦的则是某些特定的、典型的事件，例如，在被教师突然提问时，儿童是如何试图回答的。

观察数据的可靠性（信度）

我们可以通过观察得到可靠的数据，尽管有时这些数据并不总是有效的。例如，我们都同意约翰尼（Johnny）举起了他的胳膊，不管是你认为他正在提问题，还是我认为他在伸展身体。不管观察者是谁，如果对于同样的行为我们采用了相同的记录方式，我们就可以获得可靠的数据。为了实现这个目标（尽管在实际中我们可能永远无法达到完美的状态），我们必须做到以下几点：

- ❑ 使用界定清晰的、可操作的编码系统；使用这样的编码系统是为了消除可能发生的**观察者偏差**（observer bias），即一个观察者可能把某个特定的行为记入攻击行为，但另外一个观察者却没有这样做。编码系统越清晰明了，可能出现的偏差就越少。
- ❑ 通过使用实践观察，训练我们的观察者使用这个编码系统。
- ❑ 通过进行**观察者内部信度**（inter-observer reliability）检验［有时也称作**评分者信度**（inter-rater reliability）］，可以检查观察者是否真正得到了与我们的观察编码系统一致的数据。通过计算某一观察者的数据与其他观察者数据之间的相关程度就可以进行这个检查（见第12章）。表5.1的数据显示了观察者1与观察者2在对被观察者攻击行为记录上的高度相关。从数据可以看出，当一个观察者所赋予的数字是大的，另一个观察者的数字也是大的，反之亦然。不过在对F这一个体的看法上出入比较大。总体而言，两者的一致性是好的，但我们可能想知道是什么原因导致两名观察者对F的看法差异如此悬殊。

表5.1　对9名被观察者1小时内所出现的攻击行为次数记录

	被观察者行为确认								
	A	B	C	D	E	F	G	H	I
观察者1	3	4	2	7	6	3	7	2	8
观察者2	3	3	3	8	6	6	6	1	8

公开与非公开的观察

我们已经讲到，被观察者知道自己被观察和不知道自己被观察，这两种研究存在着较大的差异，这里不再介绍新的内容。第3章讲到了**需求特征**（demand characteristics）的概念，我们认为，知道自己被观察的个体可能会试图从研究场景中获得

一些关于观测假设的线索，尽管研究场景对他们来说是生活化的情境。在前面我们提到了**霍桑效应**（Hawthorne effects），涉及被试期望的其他方面同样存在着这种效应。如果被观察者知道自己正在被观察，那么这些因素中的任何一项都可能对我们的观察结果产生显著的影响。这种影响常常和被观察者的反应特征直接联系，也就是说，被观察者对于研究的**反应**可能会掩盖我们对真正发生行为的原因的进一步解释。例如，布罗迪、斯通曼和惠特利（Brody，Stoneman，and Wheatley，1984）发现，在被观察的情况下，兄弟姐妹之间的相互扭打会减少，争吵也不会那么多，他们使用威胁性的行为也更少。吉特斯汉等人（Gittelsohn et al.，1997）对尼泊尔家庭中发生的行为进行研究后发现，在他们观测的早期阶段，积极健康的行为增加了，同时，消极的社会行为减少了。对于如何避免这种反应的影响，查尔斯沃斯和哈特波（Charlesworth and Hartup，1967）曾经提供了一个经验（年代已相当久远）。他们访问了一个托儿所，和孩子们一起玩，了解他们的名字。通过这种方式，他们成为孩子所处情境中熟悉的一部分，孩子们对于观察者所观测的行为，也就较少作出反常的反应。

当观察处于非公开情形时，很显然，除非取得了被观察者的知情同意（这在观察前都已规定出来），或者告诉被观察者，观察者才可以从研究中拿走涉及他们个人的资料，否则新的伦理问题就出现了。就像在上面已介绍的那样，处理这种麻烦的一种方法就是，先记录行为，然后在其他人看到这些资料之前，征得被观察者同意是否可以使用这些记录。

观察者的参与——融入还是旁观

到目前为止，我们给出的例子都是观察者从社会场景**外部**来对他人进行观察。例如，查尔斯沃斯和哈特波（Charlesworth and Hartup，1967）访问了一些研究者，而这两人无论如何都算不上学校组织的一部分。一些研究者认为，用这种方式进行观察也存在风险，即观察者可能无法正确评估行为发生的情境，无法知道在压力下工作意味着什么，无法处理与客户的关系，等等。对于参与观察的一个强有力的支持体现在第二个引述中，这篇引文出自怀特 1943 年（Whyte，1943）在芝加哥研究的论著中，见专栏 5.2。一些研究者在社会学和人类学研究中借用了参与观察的研究方法。在某种程度上，参与式观察中的观察者加入了被观察者所在的团体，或者和他们在一起工作，这样他就可以被简单地看作是公开的观察者，因为他在协助自己所在的团体处理日常事务。这意味着，观察者在组织内部扮演了某个角色，譬如班级助理，因此孩子们能够像往常一样表现出他们的行为。然而，这也可能意味着完全的非公开参与，正如在罗森汉（Rosenhan，1973）的著名研究里所揭示的那样。在这项研究中，研究助理人员在门诊病房抱怨他们脑子里听到了说话声和噪音。在他们随后呆在精神病房的一段时间里，其他病人常常在医生面前指出他们的“正常”的表现。一些助理人员在医院呆的时间长达 50 天，“病人”们很谨慎的观察记录从另一角度可以被看做“极端的写作行为”。

专栏5.1　戒烟?——当我在这儿时并不这样

为了调查制度障碍对于病人戒烟的影响,劳恩(Lawn,2004)在一家精神病院开展了一项参与观察研究。她在两个独立的澳大利亚人聚集地实施了一项详细的观察活动,第二个地点的观察作为继续研究主要是为了使第一个地点的观察结果更有推广性。她对病人进行了随机访谈,有些要持续几个小时,并且在当场或事后作了大量的记录。观察者还与工作人员和病人进行了一些讨论,观察者自己的反思笔记也作为记录进行保留。另外,通过其他渠道所获得的一些信息也被汇总到整体的资料中。

这项研究在很大程度上是基于定性基本理论方法的,但是,在涉及被观察者关于吸烟问题的交流时,也进行了一些定量的观察。

劳恩得到了一系列令人惊奇的发现,由于内容太多,这里无法全部概述出来,其中一点强调了吸烟在病人生活中的关键作用以及对于精神病院工作人员的特殊意义。当工作人员和病人更密切地交谈和彼此更加了解时,吸烟的行为就中断了。通过允许病人抽烟,香烟常常被用作帮助控制困难行为(对工作人员来讲)的工具。香烟也被病人当做食物甚至是性的交换物。

如果劳恩仅仅使用预先设计好的调查问卷或者只是访谈,很难想象她能够描绘出在精神病院中吸烟这一行为所发挥作用的完整画面。正如事情所发生的那样,劳恩的调查不是一种机械的访谈,也不是简单地为了事先安排好的访谈随便走访。而是作为现场的一分子,深入到病房中和工作人员及病人进行详细的讨论。

从这项研究中所得到的另一点启示是,被称做参与观察的现代调查方法大多时候并不仅仅只是观察,就像这个例子那样,常常还包括访谈、结构化的观察和调查问卷的使用(尽管这个例子中没有用到)。这个领域中的其他参与观察研究不过是一系列的半结构化访谈而已。

5.5　参与观察中的伦理问题

与远距离观察相比,当观察者出现在被观察者中间时,对现场的影响显然有很大的不同。研究者应该意识到,他们的出现可能会导致人们改变平时的行为方式,正如在第三段引文里所例证的那样,这个引文来自于怀特的芝加哥研究,见专栏5.2。如果参与是非公开的,那么就像在其他任何蒙蔽被观察者的研究中一样,也会出现同样的问题。然而,如果只是因为研究要持续的时间很长,那么在长期的参与观察研究中,这种欺骗的程度是相当严重的。被观察者可能会对他认为的团体成员讲一些事情,而这些事情他绝不会讲给一个做学术观察的人。在一些好的研究中,研究者在出版研究发现之前,会跟被观察者一起对他们希望公开出版的内容进行认真梳理。他们也可能就文中的观点询问被观察者,让他们知道他们所不愿参与的研究项目的实质,并把这些内容写进报告。如果研究没有得到被观察者对出版的同意而贸然行动,就会导致公众对学术团体的不信任。

专栏5.2 怀特(Whyte,1943)的芝加哥团体研究

以写一本关于该地区的书为由,怀特(Whyte,1943)[1]研究了一个芝加哥团体中的行为。这儿是三段引文,前两段是怀特在研究之后自己写的,后一段则出自研究中的一位团体成员。

我开始是作为一个非参与的观察者。但当我被接受进入这个团体后,我发现自己正在成为一个非观察的参与者。(1943:321)

我了解了一些问题的答案,但如果我只是以访谈为基础来获得信息的话,我是如何也不会想到来问这些问题的。(1943:303)

自从你到这儿以后,你使我的生活减速了许多。现在,当我做某件事情时,我必须想想比尔·怀特(Bill Whyte)想知道什么,我怎么来解释它。在以前,我总是靠直觉做事情。(1943:301)

5.6 一些和参与观察有关的问题

- 除了涉及伦理问题之外,专栏5.2第三段引文指明,研究者很可能会改变他们所研究的对象的行为。但是,怀特认为,他在团体中待的时间越长,他的行为对本来可能发生的事件的影响就越小。
- 然而,上面第一段引文却揭示了当科学的研究者过于接近和卷入所研究对象的生活时,研究者的客观性和公正性就会受损害。
- 进行参与式观察的研究者,特别是进行非公开化的观察,在做笔记时往往会遇到麻烦。一些研究者会以某种借口来记笔记,就像怀特以秘书的角色进入意大利团体俱乐部一样。不过在大多数情况下,研究者要凭借他们的记忆来描写白天发生的事情,而你的记忆会告诉你,你对事件的回忆受到许多微妙因素的影响。另外,因为研究者是事件的唯一目击者,也就无法对其观察的可靠性进行核查。

练 习

1. 请设计一项使用观察法的研究,这项研究可以对下面的假设进行检验。
   (a)在探险的游戏中,相对女儿而言,做母亲的会允许儿子冒险走得更远。
   (b)在电梯中,人们总是试图站得与不熟悉的人尽可能的远。
   (c)相对于男人而言,女人做司机更安全。
   在每一种情形中,要保证你可以操纵变量,描述数据搜集的精确方式,包括取样方法、地点、设备和评分系统。

[1] 这里和前面的名字不一样,但推理应是一个人。

2. 结构化观察的优点和缺点是什么?
3. 一个学生想使用参与观察来研究她自己所在的学生团体。她对她的同学如何处理学习需求、考试复习和社会承诺比较感兴趣。请讨论她设计研究、搜集数据的方法,以及如何处理在研究中可能遇到的一些困难。
4. 一个研究者从两个观察者那儿得到了如下表中的结果,请就此评论观察者之间的信度水平。

以5分钟为间隔对儿童X攻击行为次数的记录

	5分钟的间隔									
	第1个	第2个	第3个	第4个	第5个	第6个	第7个	第8个	第9个	第10个
观察者A	1	3	4	2	5	12	9	4	8	9
观察者B	2	10	8	7	1	3	5	5	6	13

答 案

问题1和3的答案是开放性的。

2. **优点**:可以检验观察的信度;能够客观地重复研究。

   缺点:编码的使用窄化了对观察行为的界定;如果人们知道他们正在被观察,将会做出反常的行为。

4. 由数据看出,观察者之间的信度不好,我们可以看到在每5分钟间隔观察中,两名观察者所得到的次数上下限差异很大。如果他们较好地接受训练,并且在使用编码系统中保持一致,在每一个时间间隔里,他们所观察到的攻击行为次数应该大致相当。我们可以通过相关分析对他们之间的一致性进行测量。(见第12章)

关键术语

编码系统(coding system)
参与观察(participant obsetvtion)
事件取样(event sampling)
时点取样(point sampling)
观察者内部信度(inter-observer reliability)
系统观察(systematic observation)
评分者信度(inter-rater reliability)
时间取样(time sampling)
观察者偏差(observer bias)

6

运用提问

——问卷法、量表法、访谈法和调查法

本章内容

- 本章讲述了通过提问题来收集数据的方法。首先介绍了问卷法,要点如下:
  - ——心理学家试图借助问卷和量表来测量变量。在测量变量时,"量表"比"问卷"更受到人们的青睐。
  - ——在量表中经常使用的项目类型及如何编制出更好、更公正的题目。
  - ——编制问卷的一般原则——如何在不远离调查对象的前提下获得最好的数据。
  - ——多种类型的量表:李克特量表、语义分析量表和视觉模拟量表。
  - ——如同其他测量方法一样,量表一定要可信、有效。我们既要探寻确立信效度的方法,又要关注量表的标准化问题。
- 最后探讨了访谈法和调查法的操作程序及与之相关的一些问题,还对半结构访谈和正式访谈两种方法进行了讨论。

6.1 心理量表和问卷

前面我们已经讨论了如何运用实验及观察的方法来搜集资料，事实上，直接接近研究对象并向他们提问题也是很重要的方式。

> 那些心理学家以为他们是谁？试图对人们进行归类分组和测量。——我们都是独立的个体！

每次论及心理量表(psychological scale)，我总感觉参与测量的人们会立刻起身并告诉他们自己——“这是我最讨厌做的事”，同时还说着上面方框中的话。我们可以对上述说法置之不理——对人们的测量并不能阻止任何人把其他人当成独一无二的个体。以身高为例——每个人测量的结果不尽相同，但肯定有一部分人的身高是相同的。再把体重列进来，一些人会具有相同的身高和体重，但人数远少于只要求身高相同的人数。现在把你在考取数学普通中等教育证书(General Certificate of Secondary Education)时所获得的课程分数也加进来，这三项都相同的人就会屈指可数。……嗯，或者这三个变量也可以改为：拥有叫做 Heather 的近亲的数量、养狗的数量以及最喜欢的颜色。我们只需用有限的几种测量方法就能鉴别出非常特殊的个体，即使他可能不是独一无二的。

日常生活中我们都在使用着测量法。请看下面这条评论：

吉姆在家很健谈，但他在学校和自己的朋友交谈时却很抑制。

这位母亲使用了粗略测量的方法。她根据与吉姆在家时讲话的不同(与其他的兄弟姐妹？与他在家庭外的表现？与他父母的表现？这里并没有明确说明)，并且使用了“抑制”这一非常普通的概念来形容吉姆的异常行为。如果吉姆的抑制性很强，他的母亲将不得不采用某种方式把他与其他孩子作对比，对他在其他地方的行为与他在家里的行为作对比。我们可以向她索要评估的标准，她可能会说，“吉姆很少与人交谈，交流中眼神接触也很少，而且做事很慢。”无论你是否注意到这种标准，也不管你是否喜欢这种标准，我们几乎无时无刻都在用它评估我们身边的人，而且往往会不自觉地将我们的评估对象与其他人作对比。

我们主要采用两种方法来测量人们的行为：一是观察他们；二是向他们提问。我们在第 5 章介绍了观察法，本章主要讲述提问题的方式。人们在测量心理变量时最常用的方法就是编制所谓的**问卷**(questionnaire)。这种方法在专栏 6.1 中已经定义过。通常，我们在心理测量中所采用的不是一系列的问题而是一套陈述，这些陈述组成了一个心理量表，它需要参与者在同意到不同意之间的不同等级中作出抉择。**调查对象**(respondent)是指对问卷或量表作答的人。

问题或量表中题目的类型——开放的还是封闭的

- **开放式问题**(open-ended items)：诸如“请告诉我你对克隆人的看法”这样的题目引发的自由言论将会超出你的想象。这是收集**定性资料**(qualitative data)的一个例子——见第 7 章。这种数据集可能维持原样，也可通过**内容分析法**(content analysis)将其量化——同样见第 7 章。然而，量表或问卷的价值主要在于所获得信息的丰富性，以及调查对象从所提供的一系列固定选项中选择其中的一项

却没有使他们感到被束缚。我们几乎没有设立只有“同意”和“不同意”两个选项的情况。然而，相对于封闭式的或有固定答案的问卷或量表，开放式的回答更不易于进行个体间的比较。

- **封闭式问题**（closed items）：就是要求我们对一些熟悉的题目给出特定的答案。如果它要求我们在一定范围内作答，我们就称之为固定选择或多项选择题。比如：

我这样描述我自己： (a)内向
(b)有时内向/有时外向
(c)外向

或者

拳击是一种野蛮的活动：

非常不同意	不同意	不确定	同意	非常同意

封闭式问题的其他例子：

请在这个从1（=非常赞成）到10（=非常不赞成）的量表上指出你对在酒吧和俱乐部取缔吸烟的看法。

我在公共场所经常感到紧张　　是/否

你的小孩从什么时候开始爬行　　____个月

从上述问题可以看出：我们能够将数据量化继而进行汇总统计、比较，以及用数字的形式对假设进行检验。

专栏6.1　心理学中所使用的量表和问卷的类型

问　卷

我们可以恰当地使用问卷提出问题，而多数心理量表则不能。通常，我们在调查过程中（见下面）运用问卷来搜集像生活方式、生活习惯、对当前或特定问题的看法、休闲娱乐、道德规则、对孩子的培训技术、选举行为等这些信息。

态度量表（attitude scale）

态度量表主要是用于评价个体对一些问题相对持久的或者习惯性的反应，而不仅仅是一种看法。一个典型的例子就是测量一个人在生活中的“保守主义”倾向。“保守主义者”可能是妥协的：他们严格遵循既定的规则，赞成对已经改造过的罪犯施以惩罚，不敢冒险。量表经常使用陈述的形式（不是问题的形式）要求调查对象在同意的不同级别中作出选择，正如下面所解释的。

心理测验（psychometric tests）

人们称之为“心理测量”（mental measures），包括备受质疑的智商（IQ）测验，还有人格测验、创造性思维测验、语言能力测验、逻辑推理能力测验等。在这里我们可以对这些测验进行区分：

人格特质(personality trait)**测验**:	你通常是什么样子?
人格状态(personality state)**测验**:	你现在是什么样子?如:你目前的焦虑状态。
能力(ability)**测验**:	你通常能做什么?如:你的计算技能。
成就(achievement)**测验**:	迄今为止,你取得了哪些成就?如:你在大学测试中的表现。
能力倾向(aptitude)**测验**:	你潜在的表现,如:一般逻辑能力测验可用于评估你在计算机编程方面取得成功的可能性。

6.2 问卷危险

问卷是非常普遍的,我们随处都可以看到——粮食袋里、大街上、门缝里、电视里,甚至有的老师在某一单元或课程结束时硬塞到你手里!我们似乎一直被权威人士要求去评价自己接受的服务,或者一直遭受他们的质问。同学们在编制问卷时,往往倾向于假想自己遇到过的情况,却没想过怎样运用所收集的数据。我至少有两个很好的理由来支撑这个醒目的标题。问卷之所以危险是因为:

1. 问卷调查给我们留下这样一种印象:利用问卷能够保证所搜集的数据的科学性。
2. 当同学们完成实践任务时,问卷可能会使他们陷入困境。

> 我们要编制一份问卷用来评估人们对猎狐的态度,现在,试着和你的同学合作或者自己简要地记下你们想在问卷中提出的问题。

同学们在利用问卷完成任务时所遇到的危险是:编制问题是如此简单,但却难以确定运用所收集的数据来做什么。劣质的问卷收集劣质的数据,整理这些数据将比分析数据难得多。回想一下你是否曾经遇到过一些做问卷的困境,而这种困境也会让粗心的问卷编制者品尝苦果。

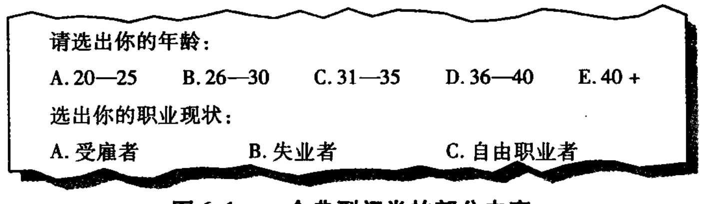
只问你需要的

人们在编制问卷时往往认为应该考虑以下几个问题,如:调查对象的年龄、性别、职业、工作情况、住处等。为什么人们会这么做呢?只有当你研究的问题与这些问题中的一个或者几个相关,你才会需要这些信息。比如说,你想了解男性与女性之间或者雇主和雇员之间是否存在差异,你就会用到这些信息。如果你不打算根据这些特征来分析,那就根本不需要这些数据。要只问你所需要的内容。

分类癖

许多人觉得有必要对问题进行分类,这样调查者就不必再提问调查对象的年龄而是只需出示一系列带有标签的表格,上面标着"25 ~ 30","31 ~ 35"等,让他们从中作出选择。为什么要这样呢?许多人认为,这是因为人们不想说出自己的确切年龄,但如果以 5 年为组距划分出各个不同的年龄段,这样很可能会大大减少询问具

体年龄给他们带来的害羞或者尴尬。注意，当你使用这些分类时，一定要避免重叠。比如，40 岁的人将如何对图 6.1 所出示的分类作出选择呢？



请选出你的年龄：

A. 20—25　B. 26—30　C. 31—35　D. 36—40　E. 40 +

选出你的职业现状：

A. 受雇者　B. 失业者　C. 自由职业者

图 6.1　一个典型问卷的部分内容

人们之所以想得到一个准确的年龄而不是一个年龄段是因为前者便于进行统计分析，本书到目前还没有涉及这一点——不过你也可能没有从头至尾地阅读。在后面我们将要讨论测量的水平。如果把年龄视作类别，我们会意外地发现，它是一种**分类变量**(categorical variable)。一般来说，我们更喜欢可以测量的变量，因为这样可以对这些变量进行更加复杂的、更有意义的统计分析。通过测量变量，我们还可以得到具体的数值，这些数值有助于我们对参与者的情况(如数学成绩、身高)进行较为准确的定位。在使用分类变量时，我们往往会止步于对不同分类中的人进行计数。下面的表格中是几个关于分类变量的例子。

典型分类变量：

- ❑ 居住类型(别墅、平房、公寓等)
- ❑ 婚姻状况(单身、已婚等)
- ❑ 有无汽车(有／无)
- ❑ 受教育水平(A 级、学位等)

分类变量有时会给我们带来不便，举例说，如果在图 6.1 中有三个调查对象来回答有关年龄的问题，你采用什么方法获得他们的平均年龄？你会认为每个人的年龄是他们所勾出的年龄范围的中间值，尽管事情可能不是你想的这样。然而，如果不这么做，你可能根本计算不出他们的平均年龄。

不完整的分类

分类的另一个弊端是可能导致类别不完整。举例说，对于图 6.1 的第二个问题，如果你是一个家庭主妇或是一位退休老人或是一名在校学生，你怎样来选择？这里并没有提供相应的分类。解决这一问题的唯一方法是提供另一个选项："其他，请具体描述"。然而，我们怎么对这种选项计分也是个问题。你可以对新的类别进行**事后比较**(post hoc)(在调查后)，不过最好在一开始就想清楚分类。在一些情况下，调查对象会因为你忽视了他们的情况而感到愤怒。

不要提那些无法回答的问题

我经常阅读草拟的问卷，为此感到只要不做调查对象，我愿做其他任何事情。试想你如何回答下面的问题：

去年，你有几次感到失落？
你每周休息几个小时？
一个月内你能感受到多少次欢乐？
在过去的两年里你拜访过几次医生？

我在草拟的问卷中甚至学生用来应付作业的问卷中见过所有这些问题，甚至还有更糟的。当然，健康的人能够非常容易地回忆出自己一年内去看了几次医生，但是有谁能够准确地估计出自己一年内失落了多少次呢？调查者需要一个准确的数字，其中的一些问题似乎是非常"科学"的，但对于上面前三个问题以数字的形式作出回答的准确程度是难以让人相信的！

侵犯个人隐私

学生不应该问像"你有犯罪记录吗"或者"你曾经患过精神病吗"这类问题，当然，大多数的考试委员会已经对此作了明确的规定。如果你不确定就看看第 13 章，在 13 章中我们谈论了种族问题和一些操作程序。

练　习

1. 问卷中含有"问题"吗？
2. 封闭式问题或项目的优点是什么？
3. 下列问卷的题目哪一个是错误的？
   (1)你一天抽多少支烟？
   ①0～10　②10～20　③20～30　④30～40　⑤>40
   (2)你所达到的最高教育水平是？
   ①普通中等教育证书　②科学院水平　③A2 级水平　④学位证
4. 你每月平均做多少个梦？

答　案

1. 如果你没有集中注意力，会以为这是一个愚蠢的问题。实事上，许多所谓的问卷并不包含问题，它们仅仅由若干个项目组成——也就是说，调查者向调查对象出示一些陈述，要求他们选择同意的等级。这就为量表能够在一定程度上测量出所测维度的细微差异提供了可能，也意味着调查对象不再被迫无奈地在"是"与"否"之间作出抉择或者说不知道如何回答了。
2. 其优点主要在于可以将答案作为数字进行计数或整理，从而可以进行统计分析。
3. (1)除了常见的重叠外(如果你每天确切地抽了 10 或 20 支烟，你将无法选择)，这里出现的另一问题是把不抽烟的人放在什么位置？(把不抽烟的群体列在每天抽 1～9 支的范围里是很荒谬的)另外，一个比较挑剔的问题是，调查对象能否理解">"这一符号的意思。
   (2)这是由于存在着比第一个学位更高的文凭(如硕士)。如果调查对象来自苏格兰，情况还会不同，更不用说来自国外了。
4. 你能严厉地要求每个人精确地对所提出的问题作出回答吗？

关键术语

开放式和封闭式问题/题目(open and closed questions/items)
调查对象(respondent)
心理量表(psychological scale)
量表题目(scale item)
问卷(questionnaire)

6.3 心理量表

我们习惯于把用来测量人们的态度、人格、智力等(见专栏6.1)的心理量表或“测验”称为**心理测量学**(psychometrics)的实践操作或简单地称为**心理测量**(psychometry)。我们把这种测验称为心理测验。那些认为心理学接近于真正科学的人将这些量表称为“工具”。比如,如果一位科学家使用某种技术对金属的电阻进行测量,这是假定所有的金属已经在相似的条件下被测定过了,全世界的科学家也都会把这种测量作为标准。对于心理测量工具,人们也试图这样做。为此,所有调查对象都应该在相似的条件下,听着同样的指导语进行测验,他们所获得的分数应该与运用所有心理学家都赞成的量表所得的分数相关。然而,对于心理学来说,这些要求是很难达到的,因为测量人不像测量金属等无生命的物体那么简单。我们后面会讨论到这个问题。首先,我们来看一下在实践作业中经常会用到的几种量表。

为什么心理量表中含有若干个项目

一般地,一种复杂的心理结构(如焦虑)包含许多方面。如果从不同的角度对其进行若干次测量,我们将会从整体上更好地了解这一心理结构。为此,一个**单一**量表中的每个项目都应该测量同一心理结构,但是仅测量这一结构的一个方面。因此,使用一个**单一的**量表只能测量一种心理结构,如焦虑、压力或自信心。许多量表包含几个分量表,从理论上讲,每一个分量表都测量一个单一的结构。

李克特态度量表

这种量表最初由李克特(Likert)在1932年创制,我们对此已经相当熟悉,而且它受到了学生们的青睐。李克特量表的编制步骤如下:

1. 编拟出一些关于态度的题目,这些题目分为赞成的(积极的)和不赞成的(消极的)两部分(且分布在这两个维度上的题目的数量是相等的),见表6.2。我们之所以编写积极和消极的题目是为了避免反应定势——它可能是偏见的一个来源。人们认为回答“是”比回答“否”更容易。除此之外,如果我们总是朝着一个方向回答,那么不管我们的回答方式如何,也无论答案可能是什么,这种答题取向会增强我们下次继续这样做的概率。我们可能不由自主地作出回答,因为我们对那一道题之前的所有问题都回答了“否”。
2. 对于每一个题目,我们都要求调查对象根据以下指标对题目进行反馈:

1	2	3	4	5
非常不同意	不同意	不确定	同意	非常同意

3. 无论什么量表,要确保高分即代表着量表测量的是什么;但从避免出现数量大而

得分低这种现象的角度来看，上述说法也许太绝对。比如，如果量表测量的是人们"对体罚的态度"，那么高分意味着非常赞成。然而，同样地，如果量表测量的是"关怀"或"不安全感"，那么高分意味着高度的关心和不安全。即使你期望那些在赞同体罚量表里得高分的人关怀感低，也不要尝试将低分作为高关怀。照着我说的做，我保证你会感激我。

4. 为了使高分表示程度高，我们需要对半数的题目进行得分转换。这正是第三步一开始就传达的意思。我们希望赞成体罚的人能得到高分。因而他们应该在所有支持他们观点的题目(如专栏6.2中的第二题)上得高分。当然，如果他们坚决不同意体罚，我们就希望他们在专栏6.2中的第一题上得高分。因此，如果个体在那个题目上得"1"分，我们就转换成"5"分。总之，在这个特定的量表中，我们要对意思为不支持体罚的所有题目进行得分转换。试着这样做，看看那些真正不同意体罚的人是什么样的情况。
5. 在对必要的题目进行反向计分后，将每个人在每道题上的最终得分相加。所得总分就是他们的态度得分。

专栏6.2 李克特量表中的题目样例

1. 无论什么原因，我们都不能打孩子。
2. 我们需要提早对孩子进行体罚以使他们远离危险。

如果你恰巧编制了这样一种量表，它可能包含15～20个题目，你现在想检验一下它的信度和效度。我们将在本书后面讨论信度和效度的概念及检验方法。要想检验内部信度，你需要把一些题目筛选出来，即个体在这些题目的得分与在其他题目上的得分不一致。比如，在体罚量表中得分低和得分高的人都可能赞同"如果父母失控，他们就会经常打孩子"这一观点。这可能是因为那些同意打孩子的人觉得只有当他们可以控制自己时才会打孩子。这种题目在量表中是没用的，因为它不能区分出赞同者和不赞同者。

语义分化量表

奥斯古德、苏茜和坦纳包姆(Osgood, Suci, and Tannenbaum, 1957)最初打算用这个量表来测量对每个人而言，客体隐含的意义。它是指这些词使我们联想到的意义而不是词典中所查到的含义(指它的外延意义)。比如，尽管我可以将婴儿的概念定义为非常小的孩子，对我来说"婴儿"的含义可能包括温暖、聪明、乐于求知，但对于其他人来说则可能意味着无休止的夜晚、金钱的花费及自由受到约束。

接受语义分化量表(semantic-differential scale)调查的对象需要在含有两极相反术语的七点量表中作出标记。对于婴儿的含义，我可以这样标记：

好的	√							坏的
弱的				√				壮的
主动的		√						被动的

等等。对于类似量表测量结果的分析，奥斯古德等人提议，不管测量什么，态度的三个要素都能将所有的两极因子联系起来。这三个要素是：

行动　包含着相反的两极，如：积极 / 消极，慢 / 快，热 / 冷

力量　包含着相反的两极，如：粗壮 / 纤弱，胖 / 瘦

评估　包含着相反的两极，如：干净 / 脏，愉快 / 不愉快

视觉模拟量表

回顾前面，你会发现我认为在分类量表中对变量进行评估比较困难。比如，我们可能会发现这样一种类别量表：

你觉得受害者会在多大程度上对施加在自己身上的罪行负责？

毫无责任　　轻微负责　　相当负责　　非常负责

在这里，我们只能根据参与者对题目的选择对他们进行分组。为了满足分类测量的需求，一种方法是使用**视觉模拟量表**(visual analogue scale)。下面来看一个例子：

你觉得受害者会在多大程度上对施加在自己身上的罪行负责？

毫无责任 | ______________________________ | 非常负责

他们不再要求参与者在表格中作出选择，而是让他们在量表中标记出一点以表示他们的位置。然后我们再对其进行测量，变量用一段距离来表示。在视觉模拟量表中，沿线的中点甚至若干个点上都代表不同的言辞。当然，你会说"我们怎么知道**你的**2厘米和**我的**2厘米是不是一样的呢？"，你确定了一个点，这意味着这量表是**主观的**。但至少我们确实拥有一套测量工具，这套工具不会强迫调查对象从一系列的类别中作出抉择。至少，我们可以运用视觉模拟量表来指明同一个体是否有所变化或者提高（如，在前后测中，或者如前面我们给出的例子，在读了不同类型的犯罪情节后）。

心理学量表中题目的类型

在用一些陈述性题目来创建一种李克特量表或者与之相似的量表时，要注意避免一些常见的差错。请看看下面专栏中提到的可能在量表中存在的题目。

有争议的量表题目

1. 如果战俘表现出改过的可能，他们就有权参与社会技能培训，除非他们最近违反条例或者表现出消极或攻击性行为，且并没有悔改之意。
2. 反对小孩子吸吮手指是种族中心主义的。
3. 不允许移民在高失业区定居。
4. 拳击手赚很多钱。
5. 非法牟利的骗子应该在监狱服刑，并且要偿还国家所给予他们的。
6. 逃避税收且免受惩罚应该是不可能的。
7. 现在的工党政府无情地破坏国民医疗服务制度。
8. 你们认为我们不应该取消学生的学费吗？
9. 你会在什么时候打孩子？

以上专栏题目存在的问题

1. **复杂性**:这种题目太长太详细,对注意力是一种挑战,为了很好地回答这一问题我们可能得阅读好几次。因此需要对题目进行简化或分成几个题目。
2. **专业术语与行话**:要使用你认为调查对象能够理解的术语。本例中他们知道什么是种族中心主义吗?如果必须使用某一术语的话,可以在导言中提供关于这一术语的解释。
3. **模糊性**:这一题目实际上是我的学生在几年前提出的。人们认为那些反对移民的人们会同意这一点。然而,那些支持移民的人或者关心调查对象的人也会同意此种观点。因为有这样一个事实:移民可能会发现在这样的地方找工作是非常困难的。
4. **事实性题目**:这实际上是另一种形式的模棱两可的题目。拳击手确实能(或者至少能)赚好多钱,这一点是毋庸置疑的,凡支持拳击和反对拳击的调查对象都同意这种观点。因此,这道题目对于测量人们对拳击的态度是没有意义的。
5. **模棱两可的题目**:假设调查对象同意让骗子偿还钱财而不同意对他们判刑,你该怎么办?这个题目同时问了两个问题,我们可以将它拆分为两个题目。
6. **双重否定**:这种题目中包含了两个否定。这样的陈述可能会令人困惑,使得调查对象陷入没必要的思考和核实。我们可以将第6条改为"逃税的人应该受到惩罚"。正如上面所看到的,对于有争议的问题或对象,我们把量表中的一部分陈述以肯定的形式来表述而另一部分采用否定的形式倒是个好主意。当学生们在练习创建量表的过程中发现这一问题时,他们经常被诱导在原有的肯定的陈述中插入"不"。我们来看一下这种做法是怎么丢失原有意义并使我们听起来感到非常奇怪的:

   原句:"我们不应该打孩子,而应该让他们在亲切的关怀中成长"。

   反向形式:"如果不打孩子,他们就不能在亲切的关怀中成长。"
7. **情绪语言**:像这样的陈述可能会使态度测验一开始就不顺利,尤其是在"新工党"选区。若遇到任何与情绪有关的题目,我们需要暂时保留这些题目,直到调查对象对调查者或测验本身感到放松后再来做。
8. 和9. **导向性问题**:人们可能不会问及第8题,但这是一个导向性问题,因为他会诱导调查对象提供一个显而易见的、期望的答案。第9题也是导向性问题,因为人们认为调查对象的确会打孩子(除非提供一个"从不"的选项)。

练　习

1. 你能看出下面的题目存在什么缺陷吗?

   (a)你认为应该废除君主政体吗?

   (b)你认为惩罚孩子的最佳方式是什么?

   (c)上个月,你有几天吹过口哨?

   (d)来自其他国家的人们和我们一样,理应得到尊重。

   (e)法律不应该宣布,在没人陪同的情况下狗不能单独游行。

(f)现在的女性也获得了高级管理层的工作。

(g)未来的性角色模式更多的应该是雌雄同体。

2. 下面的题目是学生们编制的用来测量“果断性”的量表的很少一部分。你认为有什么问题?

(a)……

(b)当你发现出了差错时,你会立即把东西送回商店吗?

(c)在一个1~10的10级量表中,相比较其他人与店员说话的声音,你与店员说话的声音有多大?

(d)……

答 案

1. (a)导向性;倾向同意。

(b)导向性;假定调查对象同意惩罚孩子。

(c)很多人可能难以作出精确的估计。

(d)模棱两可;调查对象可能同意外国人应该受到尊重,但并不意味着调查对象是一样的。

(e)双重否定;难以判断。

(f)事实;承认这个事实并不一定表示人们对女性态度的转变或女性和男性享有平等的机会。

(g)专业术语;调查对象可能不明白“雌雄同体”的意思。

2. 在量表中,用于测量同一心理结构的题目必须属于同一种类型;我们不能像这样把题目混合或配对,因为没有合适的方法把人们在量表上的得分相加。

关键术语

李克特量表(likert scale)	反应定势(response set)
心理测验(psychometric test)	语义分化(semantic differential)
心理测量学(psychometrics)	视觉模拟量表(visual analogue scale)

6.4 量表的信度、效度和标准化

假设你处在非常不利的境地,你必须向警察提供证词以说明你周六晚的行踪。在两种情形下,你的证词可能是不可信的。在叙述中,你自己的观点相互冲突或者你几天后的第二次叙述与第一次相矛盾。

当我们讨论心理测量或量表的信度时,通常是指它在相似情形中的一致性程度。关于一致性程度,我们要提到两个问题:

1. **内部一致性**:量表本身一致吗?其中的一些题目与其他的题目相关吗?
2. **外部一致性**(也叫稳定性):测验的分数会随着施测场合的不同而变化吗?

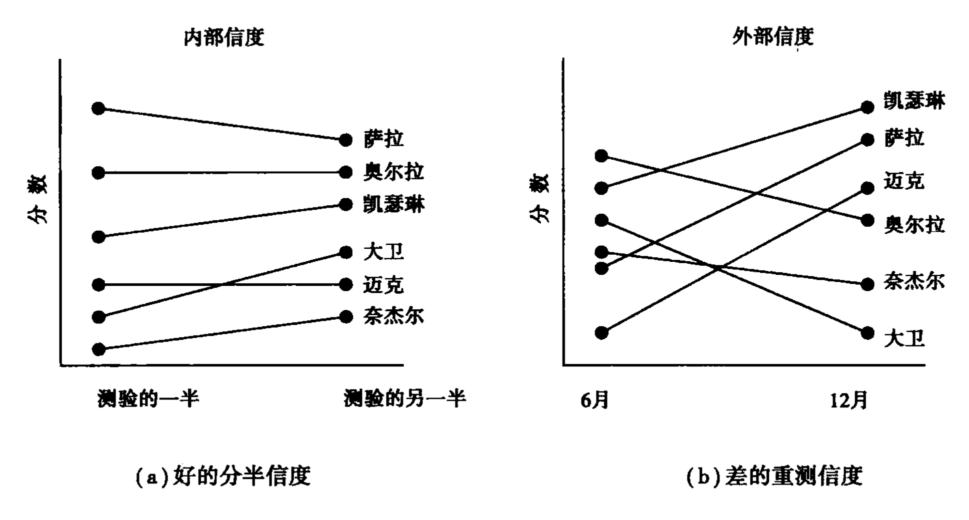
内部信度

我们可以用多种方法来检验内部信度(internal reliability)。主要是看人们在测验的某些题目上的得分是否与在另一些题目上的得分相同。

分半信度(split-half reliability)

我们可以把量表中的题目分成两半,既可以按照奇偶数将题目分成两组,也可

以随机划分。如果测验是可信的,那么人们在其中一半题目上的得分应该与在另一半题目上的得分存在相关(见第12章)。也就是说,如果一个人在其中一半测验中得了高分,那么他在另一半测验中也应该得高分,反之亦然,见图6.2(a)。



(a)好的分半信度　　(b)差的重测信度

图6.2　分半信度与重测信度(test-retest reliability)

项目分析

对于一个好的心理量表,我们希望它的题目能够区分出高分者和低分者。**项目分析**(item analsis)的方法有许多,一般将调查对象在每一个项目上的得分与在总量表上的得分相比较。通过比较,剔除鉴别力差的题目进而提高量表的信度。这里存在着两个问题。首先,这是一种基于整个量表的操作。我们将其在初始量表上的得分作为探寻次等题目的标准。如果这个量表自一开始就不可信,我们怎么还能把次等题目与它作对照呢?整个过程实际上是一个循环,却被人们普遍接受。其次,题目高度相关会提高量表的信度,但这也很可能会影响到量表的效度——接下来将会谈及这一概念。量表的效度反映了这一量表能够测量出它想要测的东西的程度。为了信度的要求而去除一些题目可能会对量表最终想要测的东西产生负面影响。

多数项目分析的方法需要复杂的统计程序,而且对于本书来说过于复杂。不过,这里有一种检测次等题目的简单而原始的方法,具体操作步骤如下:

1. 选出在某一题上得分最高的15%的被试和得分最低的15%的被试。
2. 分别将这两组被试在每一题上的分数相加。
3. 若这两组的总分相近,则说明这些题目没能很好地将这两个极端组区分开来。因此我们需要将那些总分相近的题目剔除(或简单地按照预先设定的比例剔除——如按照1/3的比例——以保留良好的题目)

研究者在创建正式的量表时会编拟大量题目,然后对这些题目进行分析筛选。接着,他们会将筛选后的题目在另一组同质的被试中试测,并持续多次。这种试测将一直延续到此量表能够产生高度一致的结果且已经通过某种形式的效度检查——接下来我们将会探讨这一概念。尽管学生们编制的量表可能达不到像态度量表那种严格水平(或与之相似),不过,将题目对一些人进行试测以发现棘手的或模

棱两可的题目或与其他题目相似而对量表没有多大价值的题目，这倒是个不错的主意。你可能在收集完数据后对它们进行项目分析，并仅仅采纳区分度最高的那些题目；这样，你将只能发现参与者在这些题目上的得分，并将这些得分作为他们在总量表上的得分。

外部信度

外部信度（external reliability）是指量表具有跨时间的稳定性。我们采用重测信度来进行检验。这需要在时间 A 对一组人进行施测获得一批分数，过一段时间（可能是几周或几个月）后对**同一组**人进行施测再获得一批分数（见图 6.2（b））。许多学生认为第二次应该对**不同的组**施测，这种说法是错误的；如果我们想知道同一个测验在**不同的时间**施测是否能得到相同的分数，我们就必须采用**同一组**被试。

效　度

简单地说，效度是指量表能否测出它想要测的东西。记得在第 3 章中我们提到过，一个实验的结果的效度是指这个结果是否就像它本来期望的那样。这两者很相似。假如我有一种令人一看就印象深刻的工具，我说它能测量人体的吸引力。当你站在一个特定个体的旁边时，我们启动机器使指示针跨越几个点，从“哦，可以”到“嗯，不错”到“挺好”再到“太棒了”，以表明你同意的程度。实际上，我刚才吹牛了，我所拥有的仅仅是一台简单的血压监测设备。如果你刚刚做过剧烈的体育运动或者你现在正处于突然的压力之中，指示盘也会以同样的点数上升。我的测量未必有效：它测量的是血压而不是对他人的吸引力。（见图 6.3）



图 6.3　这可能不是测量对课本作者的态度的有效方式

血压的例子虽然听起来有点愚蠢，但心理学中所用的许多测量是否真的测出了它们想要测的东西仍然存在着争议。一种想测量“外倾性”（extroversion）的速成量表实际上很可能会测出像“社会性”（sociability）（与大众的观念相反，这些不是一回事）等这方面的东西。评估心理量表或测验的效度的方法有很多，这里我们仅介绍一部分。

表面效度

表面效度(face validity)涉及的是测验的表面现象。在研究者及答题者看来,一种包含着测量打字速度及准确性的打字技巧测验明显测量了它要测的东西。这看上去非常明显,然而一些测验如我们所熟知的投射测验(projective test)则没有表面效度。这是因为它们是建立在弗洛伊德投射理论基础上的,其主要思想是,当人们面对模棱两可的刺激时(如一滴墨水或一张模棱两可的图片),将会投射出那些防卫性较低的内心冲突,也会泄漏试图隐藏的情感和焦虑。使用这种测验的研究者们并不想让参与者(或他们的来访者,如果他们在从事临床心理或治疗的话)清楚地了解测验的目的。(见图6.4)



图6.4 主题统觉测验中所使用的题目类型——一种投射测验

内容效度

只有保证测验中包含着所要测量领域的代表性测题,**内容效度**(content validity)才得以确立。用一个仅包含加法和乘法运算的测验来测量成人的数学能力是比较差劲的,因为数学能力所包含的内容远不止加法和乘法。为了具有内容效度,我们就得保证数学技能包括的所有重要领域都得到抽样,并在测验中有题目反映。对于这类测验,我们不难确立内容效度,因为数学专家可以告诉我们抽取代表数学技巧和数学能力的合适样本。对于那些测量更加复杂的心理结构(如:自尊、依赖或焦虑)的量表,我们需要向专家请教、查阅文献以及研究过去对这些心理结构的测量经验等,以便于对这些概念进行更透彻的研究。同样地,教师在准备试题时要覆盖所有内容而不是过度强调某一特定的问题。

效标效度

效标效度(criterion validity)是指某个测验(如焦虑测验)所得出的结果与相应的权威测验或被试的实际表现的相关程度。如果这个测验的结果与其他相关或相似测验的结果相似,我们就说它有**同时效度**(concurrent validity)。比如,人们在新的焦虑测验中所得的分数应该与其在先前的焦虑测验中所得的分数相同。因为编制

量表的研究者们可能已经断定先前的测验中有缺陷,所以我们并不期待每个人在这两个测验中的分数完全一致。然而,我们希望它们存在一定程度的相似,并且能说出新测验比旧测验好在哪里,怎么好法?预测效度(predictive validity)能够在测验分数的基础上作出预测。如果预言得以证实,就会增加测验的效度。比如,如果一个孩子在10岁时在一次数学能力测验中得了高分,我们可能会预期,他在15岁时的数学测试中同样会取得好成绩,他会在更高的水平上继续研究数学以及从事需要数学能力的工作,等等。在科学中,预言并不一定意味着对在相对遥远的未来将要发生的事情的估计。因此,同时效度可以看作是预测效度的一种形式,因为我们真正要做的是预测新量表和旧量表之间的相关程度。这种经过科学思考的预言是让我们发现哪些变量之间存在着联系,以及存在什么关系。

结构效度

结构效度(construct validity)与我们前面讲述的效度的种类不同。从最广泛的意义上看,这种形式的验证涉及了心理学中所提出的一种理论结构的全过程。心理学研究中充满了无形的变量,如"控制点"(locus of control),"自我力量"(ego-strength),"依附性"(dependency)。为使其他心理学家严格控制这些变量,每一个心理变量都必须是可见的,最好是严格遵循发展着的理论,预测变量间的关系,解释一系列已经公布了的发现,最后做到独一无二,即再也没有其他单个变量或者一组变量能更好地解释这些现象了。在物理学中,"重力"或"夸克"这两种结构已彻底得到证实,它们是解释所有观察到的现象所必须的,利用它们有助于我们预测更多东西。许多心理学家(但不是全部——见第7章)立志走同样的科学道路,因此他们也采用相同的科学程序检验他们提出的结构。

罗特(Rotter,1966)的"控制点"结构最初是由行为主义心理学家所提出的强化理论发展而来,这种强化理论包含的一些方法有助于人们理解自己所受到的强化。"控制点"能测量我们在多大程度上将自己的行为归因于自我决定和努力等内在因素(内控),在多大程度上归因于命运或运气等外在因素(外控)。这种结构与(自我)归因的认知理论相吻合,并且与许多其他的变量相关,自这个理论问世以来已经被成千上万的调查研究使用过。为了确立结构效度,我们不仅要看那些确实与之相关的结构,还要探寻与它无关的结构及理论上不应该与它有关的那些结构。从现有研究结果来看,内控与自尊或雄心等这样的心理结构正相关,与抑郁或无助等这样的心理结构**负相关**,这些结果非常有意义。也就是说,一个人越是趋向于内控,他所体验到的抑郁或无助就越少。这种发现已经不止一次得到证实,但人们也发现控制点几乎与所谓的外向性心理结构毫不相干(如:史雷伯格,1972)。所有的这些关系、一些重复研究的结果及控制点理论对新的研究结果的预测能力都使得该理论具备了更好的结构效度。

心理量表的标准化

量表的标准化涉及对量表的调整,使用信度和效度检验淘汰题目直到能有效地测量想要测的东西,同时我们也能够公正、有把握地比较个体之间的差异。为了进行这样的比较,首先要对大样本的目标群体进行施测,这样才能确立标准差和平均数(见第8章)。我们经常对测验和量表进行调整直到从大样本中得出的分数符合

近似正态分布(near-enough normal distribution)。这样就可以建立起我们感兴趣的各组的总体平均数,并且可以看到它们围绕着平均数变化的幅度。基于比较大的代表性样本,还可以估计出那些在量表上获得不同分数的人的比例。

心理测验不仅应用于理论研究,也应用于应用层面,如人们在面临升学、职业选择等人生转折和机遇需要作出选择时。因此,非常重要的一点是,这些测验不以特定的方式歧视某些人群(若带有歧视性,无论如何都会降低测验的科学价值)。所以,标准化要求测验兼具道德性和科学性。

关于标准化过程的一个非常关键的问题是,我们只能有把握地将量表应用于从相同或相似的人群中抽取的样本,而在这些人群中测验最初已经被标准化了。比如,我们不会把控制点当成一种通用的结构。也就是说,我们并不认为那些理解测验的人们在已被翻译的控制点测验中的得分与来自完全不同文化的人们的得分具有可比性。这就是我们在第3章中简单提及的**总体效度**(population validity)。几位研究者已经发现,对于一些文化,罗特的量表会产生两个(甚至更多个)不同的维度,一种是关于个人对日常生活事件控制的,另一种则涉及对更广泛的社会事件或行政事件的控制(如 Smith, Trompenaars and Dugan, 1995)。

练 习

1. 我的同学设计了一份关于"对英国的态度"的测验,想对一批将要离开英国到海外的学生施测。(a)我们能检验这个测验的信度吗?(b)怎样检验这个测验的效度?
2. 我的朋友说"我的狗喜欢贝多芬。每当我演奏贝多芬的音乐时,它就会走过来卷起尾巴靠在我的膝前"。这是一个可信的测验,还是有效地测验,或者两者都不是呢?
3. 我的另一个朋友说"我在国内采访一所大学时曾做过一个测验,结果显示我在'个性化'和'社会性'上的得分非常低。我想知道这意味着什么?"就你朋友的看法进行评论。

答 案

1. (a)能,但只有内部信度。(b)可以通过将这些人的分数与想回国的人的分数作对比来检验其效度,也可以对那些学生进行个别访谈(separate interview),再来进行比较。
2. 可信但不一定有效!
3. 她正在把自己与人们对一个美国学生的要求准则作对比。鉴于以前的研究,我们有充足的理由相信美国学生在这两个变量上获得更高的平均分。她忘记了测验的标准化和文化平等问题。

关键术语

投射测验(projective tests)	效度(validity)
信度(reliability)	表面(face)
内在的(internal)	内容(content)
分半(split-half)	标准(criterion)
项目分析(item analysis)	同时(concurrent)
外在的(external)	预测(predictive)
标准化(standardization)	结构(construct)

6.5 访谈法

这里先回顾一下前面所讲的问卷的结构与内容。一项心理学研究可能仅涉及对量表或问卷的整理,心理学家没必要与每一个调查对象接触。比如说,问卷可以由卫生所或邮局发放。然而,用访谈法收集数据要求访谈者(interviewer)和访谈对象(interviewee)之间必须有直接接触;尽管主要是面对面接触,但没必要总是这样。访谈法是指通过涉及研究成员和参与者的某种直接接触并以某种形式的提问来收集数据。

专栏6.3 访谈法与问卷法的比较

基于访谈的研究和基于问卷的研究的一个明显的区别是:在访谈情形中,人与人之间的关系这一变量可能会影响到访谈对象行为的真实性。参与者可能被诱导往"看起来好"的(社会赞许的)方面回答,同时"活生生"的访谈者的出现也可能以多种方式影响他们的回答。思考一下是否还有其他的影响因素。

- **性别**(gender)。研究支持这一观点,人们对异性访谈者的反应更为积极(Rosnow and Rosenthal, 1997)。访谈者的性别与访谈对象的性别是相互影响的。威尔逊等人(Wilson et al, 2002)发现加州的拉丁美裔男性在面对女性访谈者时比面对男性访谈者报告更少的性冲突和性伴侣。
- **种族**(ethnicity)。研究表明,访谈对象与访谈者属于不同的种族时,他们的回答更为正式(Zanna and Cooper, 1974)。人们还发现,如果访谈对象和访谈者属于同一个种族,访谈者对访谈对象的评价更为积极(Awosunle and Doyle, 2001)。
- **正式角色**(formal roles)。不管访谈者有多放松,对于那些要作出相应反馈的大多数访谈对象来说,他们都是重要的人物。比如,调查对象很可能搜寻更"正确"的语言来回答访谈者的问题。
- **人格**(personality)。两个人之间的"化学反应"可能对真实的、完善的研究资料的收集起到建设性或破坏性的作用。
- **暗示性评价**(evaluative cues)。在研究的背景下,人们往往想知道访谈者需要自己作出什么样的回应(想成为一名"好"的参与者)。因此,访谈者要避免对那些听起来带有批判性或不令人非常满意的答案作出反应。

另一方面,各种类型的访谈都具有各自明显的优点。然而,我们可以根据接下来描述的多种类型的访谈程序对这些优点进行评论。一般而言,一个"活生生"的访谈者能够纠正调查对象对所提问题的任何误解;这不同于使用问卷法的情形。他们可以改写问题或给出具体的例子,他们还可以使调查对象感到更加放松并确信自己做得很出色。有时访谈者会获得令人出乎意料的信息,尤其在定性研究中,资料收集的过程中经常会发生这样的情况——见第7章。

访谈的结构

在上一章,我们提到了用来收集资料的观察法的结构的不同维度。这里我们来

看一下访谈法的结构：访谈法从高度结构化的到结构松散的都有。当研究被称为是“以访谈为基础”时，研究者采用的往往是一种被称之为半结构访谈（semi-structure interview）的方法。专栏6.4描述了访谈法的各种类型及具体应用。

专栏6.4　心理学研究中访谈法的类型

非定向/非正式（non-directive/informal）访谈　非定向访谈很少用来搜集信息，它主要用于协助来访者顺利完成心理治疗或进行咨询服务。这种访谈的思想是，绝不能用已经设定好的问题把来访者束缚住，应该让他们在没有限定方向的情况下尽可能自由地发挥，访谈者的作用主要在于对来访者的想法给予反馈与重组。

然而，这种类型的访谈在心理学的研究中也曾被使用过，就像根据库利坎（Coolicon，2004：152）所描述的那样，霍桑研究就是一个例子。其思想是尽可能地获取信息，不仅从自己的言语中获取，也要从他们组织自己言语的方式中获取，但不要从访谈者指导他们的方式中获取。定性研究者（见第7章）经常使用这种方法。他们的目标主要不是搜集能够用于统计分析和支持量化假设的数据，而是搜集那些尽可能与访谈对象自己的意图和观点相贴近的、丰富的定性资料。

半结构访谈　在半结构访谈中，可以保持非正式访谈的优势（以至于使访谈尽可能像“聊天”），同时还应尽力保证每次访谈都有同样的话题，这样来自两个访谈的信息可以作对比和（或）以一种相对标准的方式将两个访谈内容合并。半结构访谈常用于定性研究，定量研究也经常用这种方法来获取可以进行编码和比较的信息，并将这些信息作为总体变化趋势的一个样本进而作为检验假设的证据。

访谈进度表中含有预先设定好的话题和一些有特定措辞的问题，但是访谈在访谈对象回答得不完整或模棱两可时可能离题。访谈者要尽可能自然地引进访谈进度表中的问题。皮亚杰（Piaget，1936）在他的研究中使用过一种叫做**临床法**（clinical method）的方法，这属于我们刚才所描述的半结构访谈。在访谈中，他会以几种不同的方式提问每个孩子一系列特定的问题，以确定每个孩子是否真正理解了一个特定的概念。

正式和完全（formal and fully）访谈　访谈者以一种固定的顺序提出问题，任何离题的谈话都不能作为访谈的正式回答进行记录。但正式访谈的题目可能仍然是开放式的。因为如果访谈法不使用开放式题目，它的实施过程就相当于问卷法，只不过采用了面对面的形式而已。

支持和反对

无论是结构松散的还是高度结构化的访谈都各有利弊。这个话题基本上类似于结构观察与非结构观察的比较。高度结构化访谈的优势在于，能够得到可进行信度检验的结果和能用传统的统计方法进行分析的数据。我们可能会认为这种方法更加“客观”和科学化。然而，定性研究者（见第7章）可能认为正式访谈的刻板性和正式化只会带来和使用实验法相类似的问题，因为参与者在正式访谈中可能感到不自在，他们可能不按常规的、真实的方式作答，而试图猜测问题的要点并向着更易为社会所赞许的方面回答。由于这些原因，人们可能对结构访谈存在某种偏见。这是因为与之相比，人们认为在半结构中访谈者会感觉更加轻松，也更容易在访谈中发挥影响。在表6.1中，我们分别对支持或反对结构访谈和半结构访谈的观点作了总结。

访谈法在研究中的作用

访谈实录是严格按照参与者所说的全部内容逐字逐句记录的，有时连咳嗽、结

巴和重复也记录在内。实录的过程——从口述、录音到转写为书面文字——会花费很长时间，十分钟的录音内容全部转写成文字要用两小时的时间（所以不要仓促地选择访谈法来完成你的方案！）。访谈法更多地为定性研究者所用，下一章将会讨论这种方法在定性资料中的使用。在定量研究中使用访谈法时，研究者可能要求培训过的**评分人**评定访谈记录的类别，如在“热心地对待孩子”的研究中，我们已经采访过母亲对孩子的抚养过程。内容分析可能用于提取一组分类或数一下已经决定了的特征的频次——比如，在提供支持或提出问题时，调查对象的家庭被提及过多少次？因此，一项访谈可以把访谈对象所报告的任何一次描述作为它的发现，也可以将它用于常规的假设检验的研究中。

6.6　调查法

调查法（surveys）实际是向人们提出相当多的问题。如果向一整群人发问，我们称之为“人口普查”。多数调查包含一组访谈，尤其是当要和每一个调查对象进行面对面的交谈时。然而，调查法的一个优点是，只要选定了发放问卷的有效方式，就会涉及大量人群。因此，选择样本的方法很重要，因为我们在面对较大样本时，经常满怀信心地认为所记录的态度和行为能够代表整个人群。如，职业父母在工作日期间会采用不同的方法处理 5 岁以下孩子的照料问题，我们可能想知道这样的职业父母占多大比例。如果挑选太多那些在家中还有其他家属的职业父母做样本，或者如果我们仅在工作日去敲他们的门的话，最终的结果将会被扭曲。调查法比其他的方法更可能使用第 2 章所讲述的分层抽样、配额抽样和整群抽样这些复杂方法。

调查取样

调查取样（survey sampling）取决于和调查对象进行交流的方式。我们可以选取一些样本，采用面对面访谈、邮寄、打电话或发邮件的方式对他们进行跟踪调查。或者，我们还可以通过在医生的诊疗室、学校等公共场所甚至在网上发放问卷（将自愿归还）的方式来获得完全自我选择式的样本。

下面列举了一系列取样方法，想想在用这些方法从一个区域中抽取代表性样本时可能存在的问题。导致样本偏差的因素有哪些？采用哪一种方法能使抽样偏差变得最小？

- ❑ 使用电话号码簿。
- ❑ 选取住宅样本。
- ❑ 在因特网上发布问卷。
- ❑ 使用选民名册。
- ❑ 向停留在大街上的人提问。

希望你会利用选民手册作为抽样偏差最小的方法，即使手册中不包括战俘、无家可归的人、接受精神治疗的人等。电话号码簿排除了那些没有电话的人。利用住宅抽样排除了不居住在住宅里的人，包括住在公寓或医院中的人等。利用因特网则在一定程度上易于抽取那些比一般人略富裕一点的人群（Backer, Wagner and Sing-

er, 2003)。在大街上抽样并把他们作为代表性样本是一种天真的想法,这一点我们在第 2 章已经讨论过。

调查法在研究中的运用

我们可以以描述性的方式来使用调查法,只是找出我们广袤的世界正在发生什么事即可,正如现在声名狼藉的金赛报告一样(Kinsey et al, 1953),它在关于性活动方面的统计数据震惊了美国人(报告声称,在美国,无论男女,约 15% 的人有外遇,10% 的男性是同性恋,15% 的女性从来没有体验过性高潮)。专栏 6.5 中的调查研究报告显示假设可以得到验证,在这种情况下,早期由气质所引起的问题可能与后来的肥胖有关。

专栏 6.5　一个调查研究假设检验的实例

帕尔克-赖贝克等人(Pulkki-Raback et al, 2005)开展了一项涉及 619 名男性和女性的追踪研究,第一次是在他们 6 岁到 12 岁时测他们的气质,第二次是在他们 24 岁到 30 岁时测量他们的体型(包括体重指数)(body mass index)。研究发现,即使其他已知的由肥胖带来的风险在分析数据时已经被控制住,人们在孩童时期表现出的气质性困难,尤其是高消极情绪性仍然与 18 年后的体重指数高度相关。此项研究仅是下面这项高度复杂的研究的一部分,这项研究涉及对 3 500 多名芬兰学生 24 年的追踪调查且一直延续至今。

练　习

1. 一位研究者先对参与者进行访谈,然后仔细检查每份访谈记录并对每位访谈对象的"热心"和"开放"程度用量表进行等级评定。这一程序有什么缺陷?
2. 你准备对当地一家超市的经理进行访谈。他 43 岁,据说待人友好且积极参与当地恢复猎狐的活动。鉴于你自己的人格和性格特征,你认为哪些因素(不管是积极的还是消极的)会影响到访谈资料的丰富性与真实性?
3. 将结构访谈与半结构访谈或非正式访谈作比较,列举出结构访谈的优点与不足。

答　案

1. 因为这位研究者已经和每位参与者都有了一次访谈经历,一些期望因素可能会使她对参与者产生偏见。解决这个问题的方法是录用一位对访谈对象没有任何先前印象的"第三者"来对访谈记录进行评估。
2. 你可以自己作决定,但要考虑到自己的态度、年龄和性别是潜在的影响因素。如果你们年龄相仿、性别相同,仍然要考虑,当一个与你截然不同的访谈者在进行这个访谈时,访谈会在何种程度上产生不同。
3. 见表 6.1。

关键术语

人口普查(census)	半结构访谈(semi-stuctured interview)
临床法(clinical method)	结构访谈(structured interview)
访谈法(interview method)	调查法(survey)
评分者(rater)	

表 6.1　结构访谈和半结构访谈的比较

优　点	缺　点
半结构/非正式访谈	
❑ 访谈对象能更加轻松、更加自由地按照自己的意愿作出回应,因而能够产生更加真实和丰富的资料	❑ 不同访谈者之间及同一个访谈者在不同场合下的访谈难以比较,因此,信度是一个更加严重的问题
❑ 访谈对象可能会遭受较少的评价焦虑	❑ 访谈者很可能脱离原定的题目,从而影响资料的收集或允许访谈对象漫谈
❑ 访谈者可以更加灵活地提问,根据访谈对象的表现探寻他们的想法或者允许他们沿着自己的思路回答	❑ 访谈者可能需要做更多方法上的准备,以避免出现偏差,并且访谈者获取最详尽的资料
	❑ 难以对资料进行定量的比较。如果进行定性分析,在第7章中讨论过的主观性问题就会非常明显
结构访谈	
❑ 易于被其他人重复;	❑ 非常正式的访谈可能引发更多的评价焦虑,也导致访谈对象努力做一个"优秀的"参与者,从而表现出不大真实的行为
❑ 多个访谈者更有可能遵循大致相同的程序进行访谈	❑ 过度僵硬的提问方式和次序限制了访谈对象充分而真实地表达自己思想的自由
❑ 访谈者需要较少的培训和专业技能	
❑ 易于对资料进行定量分析和信度检验	

7

定性资料与定性研究方法

本章内容

- ❑ 定性资料的性质及使用定性方法对心理学研究的意义。
- ❑ 心理科学中反对实证论和数字测量的主要观点。
- ❑ 定性研究者的建议。
- ❑ 定性分析的性质及确保定性研究可信、有效的措施。
- ❑ 介绍当代几种主要的定性方法,其中包括内容分析,运用这些方法能从定性资料中获得定量资料。简要描述扎根理论、诠释现象学分析、话语分析和主题(理论导向)分析等研究方法。
- ❑ 个案研究的性质和运用。

7.1 定性与定量的方法和数据

从上一章的访谈资料讨论中我们了解到,如果访谈中不包括那些能让被试详尽回答的开放式问题,那么这个访谈几乎是不值得做。访谈对象的回答最终以文本形式出现——被访者所讲内容的真实记录。由此导出现在的问题——如何处理资料?一般来说,研究者根据**内容分析法**(content analysis)(将在后面讲述)计算访谈对象所提到的主题或概念的次数。除此之外常用的方法,诸如上一章谈到的,不依赖评定者对记录的编码,而是根据定义好的变量给每个被试一个分数的方法。通过这种方法,访谈数据将被量化,即对变量进行数字评价。

用这种方法量化定性数据(qualitative data)的不足之处是,访谈对象提供的所有其他重要信息都丢失了,以至于我们对被试所讲的内容缺少整体印象,就如没有鳍的鱼,或仅仅告诉某个人一幅画中使用的颜色和每种颜色的用量。本章涉及两个密切联系的问题:

1. **定性资料**
2. **定性研究方法**

定性数据是所有非数字形式的信息,通常是文本(如某人口头的原始报告)。一般来说,定性数据的类型主要包括:

- 文本(例如访谈记录、日记记录、焦点组的交谈、来访者反馈书中的评论)。
- 观察(例如,观察者的记录)。
- 图画资料(例如,图画、儿童的绘画)。
- 听觉资料(例如,对音乐的选择)。

我们选取被试的一篇文章(如简短叙述,访谈记录),通过计算项目[①]出现的频率或者让评定者给变量如"热情"打分。这样我们就从原始的定性数据中创建了"定量数据"(quantitative data)。事实上,许多经典或权威的研究中业已使用过不需要数量化的定性信息。如弗洛伊德(Freud)的资料几乎完全是定性的,虽然有人不把他看做"科学家"。另一个典型例子是科学权威潘菲尔德和拉斯马森(Penfield and Rasmussen, 1950)的研究,虽然他们的研究披着医学这一科学外衣,但令心理学家兴奋的却是电击患者大脑皮层时,患者提供的那些极详细和生动的记忆报告。在罗森汉(Rosenhan, 1973)的研究中,如果没有几个美国精神病学科(US psychiatric department)伪造病人的定性信息报告,这一研究也就没多大趣味。在华生和雷纳(Watson and Rayner, 1920)的小艾伯特(Little Albert)形成条件反射过程的文章中,很多内容是对艾伯特反应的定性观察的实验日记摘录,几乎没有呈现定量数据。(见 http://psychclassics.yorku.ca/Watson/emotion.htm)。

虽然心理学家常常把定性数据作为研究证据公开发布,但主要还是采用定量方法进行研究。在过去的20年里,人们越来越多地反对心理学研究的"硬"科学性和定量性质。这种反应是对**实证主义**(positivism)的极大挑战。实证主义是影响19世

---

① 文章的主题或概念。——译者注

纪科学长足发展的一门哲学，它提出不能被观察的和不能用数字测量的事物对科学研究来讲是不可用的。我们看一下由定性研究者提出的反实证主义观点——不过，应该强调一下，虽然定性研究现在还没有统一的学派，但是有很多定性方法。即便这样，大多数定性研究者可能仍然会赞同以下观点：

1. 实证研究将人们从他们的环境背景中孤立出来，甚至把人的"部分"（如，他们的记忆或自我概念）视为可分离的。参与者（通常称为"被试"）被视为从调查的变量中分离出来的相同单元。研究者对这些变量持有先入为主的观点，被试却不能提出不同的意见。实际上，被试是被控制和被测量的变量。
2. 主试为了保证实验的科学客观性，力求与被试保持分离，不影响被试的反应。然而，实验情境毕竟是社会性的情境，无论主试是否意识到这一点，被试总是以社会建构的方式对主试作出反应。当研究者想和被试说再见并急于给他们的心理量表评分时，被试则常常想讨论或争论量表中项目的确定性。
3. 人为控制的实验方法仅能收集到表面数据，因为被试缺少正常反应、计划和考虑整个情景意义的自由。不过在这些情况下收集的数据被视为是现实的、可推广的。以这种方式产生的个体模式是简单机械的。举个具有批判性的例子，如哈雷（Harré, 1981）的实验，其目的在于表明自我感知（sense of self）的增长是否可能引起利他行为的增加。可行的实验：让女被试通过录像监视器观察自己 1 分钟，以增加她们的自我感知，然后请她们听一个关于性病（venereal disease）的讲座，1 分钟或 4 分钟过后，要求她们为性病治疗方案作出贡献。哈雷认为这种定量方法完全否定了 G. H. 米德（Mead, 1934）自我概念研究的理论基础。①
4. 正如上述研究和心理量表中所示，严格操作变量，给被试强加预定的测量，这种研究，使得最重要的和有价值的信息大量丢失。
5. 实验研究经常具有欺骗性，主试在其中的地位优于被试。所以只要有欺骗存在，主试与被试间的关系即便不是侮辱性的，也是居高临下的。②

定性研究者的建议

虽然目前有多种多样的定性方法可供研究，但是大多数定性研究者可能一致同意下列定性研究的优点及其规则：

1. 心理学研究应该关注行为在社会背景中的意义，而不是孤立的、客观的行为单元——采取"整体的"，而非"原子的"方法。
2. 研究应该尽可能自然地收集定性数据。
3. 操作应尽可能由被试进行，而不是施加于被试。在研究关系中研究者自身的角色是被认识到的，因此许多方法要求**反思**（reflexive）说明[或者**反思**（reflexivity）]，即研究者在写报告时，要考虑他们自身对被试的反应及数据解释可能产生的影响。
4. 强调被试自身方面的问题及其对周围环境的解释，而不是过多强调假设检验。通

---

① 根据自我概念严格设定符合要求的定量方法，而研究结果却不符合此概念的假设。——译者注

② 在实验中，主试为了隐藏实验的真实目的，给被试的指导语与真正目的无关，这种欺骗，如果不是对被试的侮辱，那么可以说是居高临下的，因为这样做似乎是合情合理的。——译者注

常，我们希望理论根植于资料。有代表性的是研究者通过彻底分析他们的访谈和观察数据，在记录中寻找一些主题和概念，这些有助于他们创建“局部”理论（即不可推广到整个人群）——但在这些“局部”理论中至少有一个可能发展为运用更广泛的理论。

定性研究的数据来源

定性研究者倾向于使用下面的主要方法或者技术，其中大部分我们在前面章节中已作了讨论：

- ❑ 开放式问卷
- ❑ 半结构访谈
- ❑ 定性参与观察
- ❑ 日记法
- ❑ 角色扮演和模拟
- ❑ 个案研究（见本章末）

访谈步骤的改变往往取决于所用的特定定性方法的规则。日记法（diary method）既包括研究者记录的日记，例如以观察者的身份参与到诸如医院或者住宅区这样的真实情境中，也包括让被试在一段时期内坚持写日记。琼斯和弗莱克特彻（Jones and Fletcher，1992）要求几对夫妇将每天的心情、压力和睡眠变化写成日记并持续记录3周，结果是职业压力会从一个身上迁移到其合作者身上。角色扮演和模拟（role-play and simulation）的方法在研究中用的相当少，但是这种方法能收集到被试扮演或模拟角色时的感受的定性数据。个案研究在下面讲述。

7.2 处理定性数据——定性分析

大多数用“定性方法”收集的数据只能在质量水平上进行分析。也就是说，以文本形式（主要地）保留的意义不能转换成数字。如果你自己对定性分析感兴趣或甚至主动要求使用定性分析，那么需要一本更高深的书。专栏7.1为你提供了从哪里可以找到进行定性分析的详细建议。在定量研究中，你可以设计一个方案，找到其薄弱点，收集数据，用恰当的统计技术进行分析——见8到12章，没必要决定赞同哪门数量哲学。因为有一门主要的实证哲学，它的使用“规则”[①]几乎是通用的。然而，定性方法由于种类繁多，所以没有形成与定量方法对应的统一领域。大多数定性研究者不反对统计分析（一些反对）。他们必须使用定性方法，是因为对于他们从事的工作种类和努力达到的目标来讲，这是最适合的方法，可以选择几种主要的方法，它们是从社会学与人类学等相关学科发展而来的。在如何获取知识的一致观点下，这些方法不是简单意义上的不同，不同之处在于知识的本质以及最恰当的获取方式上。实证论站在现实主义的立场上——其观点为：科学家一定能发现世界上存在的单一具体的现实，知识是发现的最佳途径。而大多数定性方法站在建构的立场上——其观点为：知识是**相对的**，一个人看到的客观存在不同于另一个人看到的，且两个不同的观点中，没有一种更有证据说明自己是正确的。知识就是我们从探究世界使用的各种语言和概念中建构出来的。例如，一个人的“恐怖主义者”是

① 实证哲学规则的使用。——译者注

另一个人的“自由解放者”或“殉难者”。

简要介绍了定性研究的领域，现在我们看一下收集、分析定性资料的一般原则。如果不是全部，那么至少大多数方法是这样的。

- 总的来说，分析必须彻底且是学术性的，它以远高于普通新闻报道或个人推测的形式反映资料问题。与定量研究相比，定性研究似乎是一种更容易的选择，但不要被这点吸引而做定性研究。因为你不可以像写杂志文章那样来写。
- 分析时必须看到研究者在数据解释中扮演了一个关键角色。应该充分认识反思性的本质，即著者要清楚他们自己的立场观点在哪些方面影响了资料诠释的建构。
- 分析时必须与被试原有的意思紧密联系，可用记录中大量的直接引文为依据，这些引文通常是支持某种主张或概念的。
- 可靠性不能通过统计方法获得，通过心理量表收集的数据并不能获得可靠性。定性研究中，不同的研究者和作者相信有几种方法可以评估其成果的可靠性和真实性：(见专栏 7.1 所列资源)

  ——**三角校正**(triangulation)——从勘测中引进的术语，表示研究者从几种不同的角度来比较同一现象。例如，访谈可用于支持由观察得出的解释。

  ——**拟合度**(fit)——在阅读定性报告时，应该让读者清楚地看到从研究者的解释回溯到原始资料的过程。这里谈谈“一致性”问题，史密斯(Smith,2003，见专栏7.1)提出“审核”的观点，后来其同事对此作了验证，他将一系列证据追根溯源，并证实了这些证据是合乎逻辑和可信的。

  ——**被试核实**(participant verification)——一些研究者确信最终的解释让原来的被试看到的话，他们能对信息的使用方式提出重要的且具有启发性的意见。

  ——**饱和与消极案例分析**(saturation and negative-case analysis)——饱和指的是收集和分析资料持续到没有新的重要概念出现为止。消极案例分析指的是，如果发现案例不符合建构的解释，那么应该继续分析直到被某种理论解释。举个不成熟的例子，即研究人们参加肥胖病自助组的动机。收集资料后从人们参加的原因、自我概念和身材等不同角度进行分析，直到没有新的主题出现为止。如果其中至少有一位参与者说他参加的唯一理由是为了获得社会联系，那么需要更改所有的动机描述。因为他们可能隐藏了一些真实的理由，或者他们完全满意自己的身材仅仅是为了寻求社会联系。

专栏7.1　供定性方法进一步研究的资源(全部细目见参考部分)

对定性方法进行充分讨论的文章：

Coolican(2004) *Research Methods and Statistics.*

Giles(2002) *Advanced Research Methods in Psychology.*

Robson(2002) *Real World Research.*

关于如何进行定性研究的文章：

Willig(2001) *Introducing Qualitative Research in Psychology.*

Smith (2003) *Qualitative Psychology: a Practical Guide to Research Methods.*

Hayes(1997) *Doing Qualitative Analysis in Psychology.*

定性研究评估指南：

Elliott, Fischer and Rennie(1999) Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, 38, 215-29 (适用于临床心理学领域).

Henwood and Pidgeon(1992) Qualitative Research and psychological theorizing. *British Journal of Psychology*, 83, 97-111(一般指南).

Yardley(2000) Dilemmas in qualitative health research. *Psychology and Health*, 15, 215-28 (适用于健康心理学范围).

可尝试的两篇简易可读的定性文章：

Abrahamsson et al. (2002) Ambivalence in coping with dental fear and avoidance: a qualitative study. *Journal of Health Psychology*, 7, 653-64.

Lawn(2004) Systemic barriers to quitting smoking among institutionalised public mental health service populations: a comparison of two Australian sites. *International Journal of Social Psychiatry*, 50, 204-15.

练　习

1. 指出使用、分析定性资料的优点与可能存在的不足。
2. 思考一些你学过的与收集定性资料有关的研究。
3. 为什么定性研究者要考虑反思性？

答　案

1. 见表7.1。
2. 虽然不是普遍的关注，但是米尔格拉姆(Milgram, 1963)和奥奇(Aach, 1956)两个人在实验之后都进行了访谈，并讨论了被试一致服从的原因或者不服从的理由。安斯沃思、贝尔和斯泰顿(Ainsworth, Bell and Stayton, 1971)的依恋研究是让藏在暗处的被试一边观察婴儿的所有行为一边进行描述录音。想知道更多的例子——问你的指导者。
3. 看了这篇或其他文章，把“反思”这个词输入到因特网的搜索引擎中，将会出现几种有用的讨论和解释。基本上，定性研究者认为传统科学产生的中立客观印象是错误的。反思的原则对在科学汇报中把研究者的影响轻描淡写这样的方式提出了质疑。值得思考的典型“诡计”是非人称代词的运用：“人们能看到提出的支持遗传的证据的优势”而不是“你能看到我提出的支持遗传的证据的优势”，考虑其强调的不同。

关键术语

拟合度(fit)	定性研究方法(qualitative research approach)
消极案例分析(negative-case analysis)	定量数据(quantitative data)
被试核实(participant verification)	反思性(reflexivity)
实证论,实证主义(positivism)	饱和(saturation)
定性数据(qualitative data)	三角校正(triangulation)

7.3 建立定性方法

由于版面所限,这里不能详细阐述流行于研究者间的各种定性方法。不过,这里可以提供一个概览,如果你对其中的内容感兴趣,可以从专栏7.1列出的专用定性文献中至少选择一篇进行深入研究,其中史密斯(Smith, 2003)的文章值得推荐。

表7.1 定量和定性数据的优缺点

优点	缺点
定量数据	
❑ 能做统计分析	❑ 不能提供个体及其想法的整体印象
❑ 能清楚地看到有代表性的分数和范围	❑ 将狭隘定义的变量视为可从个体和背景的其余部分中分离
❑ 能用来检验明确的假设	❑ 可能对不容置疑的科学发现产生错误的表达
❑ 能从样本推论到总体	
定性数据	
❑ 保留了个人所有的原有意思	❑ 很难将结果推论到其他的情况
❑ 如果材料收集得当,则是丰富且真实的	❑ 对收集和分析数据的恰当方式存在不一致
❑ 能呈现出个人对特定主题的整体观点和印象	❑ 研究者的观点和偏见更多地影响分析与解释

扎根理论(GT)

扎根理论(grounded theory)最初由格拉泽和斯特劳斯(Glaser and Strauss, 1967)从社会学中引入。它的原本之意是,理论产生的唯一来源必须是收集的数据,不能受事先形成的任何观点或早先理论的影响。

但这很难做到,大多数研究者会将他们的目的与先前的研究和成果联系起来。对数据进行彻底地分析(直到不能再解释为止),第一阶段是从数据中选取特定的主题或者"种类",进一步分析要求发展更高级的种类,必要时再进行带有目的性的抽样,以收集更多的数据,这样做是为了检验新兴的理论和剔除异常值。解释所有种类的最终框架实际上是解释模型。斯特劳斯和科尔宾(Strauss and Corbin, 1990)发展了扎根理论运用指南,但遭到了格拉泽(Glaser, 1992)的驳斥;在这种方法中,甚至对研究应该如何实施还存在几种分歧。专栏7.1底部引用的两篇简易可读的文章运用了GT理论。

诠释现象学分析(IPA)

IPA(Interpretive Phenomenological Analysis)不是一种英国啤酒,而是一个复杂累赘的名称的缩略词,这种方法试图尽量与个体的经历——即他们的"现象学"保持密切的联系。IPA 沿袭了卡尔·罗杰斯(Carl Rogers, 1961)的观点。罗杰斯认为,无论一个人报告的经历如何稀奇古怪,那是他们的经历,我们不能证明它无效。IPA 根据反思性原则试图表达真正的经历——没有亲身体验,没有人能表达另一个人的经历,它的分析与 GT 相似,不同之处在于它努力提取经历的主要特点时允许(不像 GT)不相关信息的缺失。在调查中,可以比较几个人的经历,然后将主题综合起来,对现象形成更宽泛的印象(如正被欺负的经历)。一个好的实际例子详见史密斯(Smith, 2003)的研究。

语义分析(DA)

很难完整描述 DA 的目标和步骤,尤其在使用 DA 的过程中存有几种不同意见和相当尖锐的分歧时。DA 的主要规则是不以语义分析(discourse analysis)作为发现人类思想的方式。在心理学中 DA 被认为是毫无结果的操作。DA 理论者像行为主义者一样认为我们所知道的心理事件是人们谈话时的结构,心理过程不存在是否被研究,而是我们不能够触及它们。他们认为(哲学地)将心理活动看做"事物"是错误的处理方式。DA 理论家和研究者关注的是人们通过谈话如何建构他们的经历和动机。举个不成熟的例子,为了给出正确的信息而不产生错误的表达,政治家和代表在众目睽睽下的演讲和访谈是非常认真的。某种意义上,我们可以回溯到选择谈"恐怖主义者"而非"自由解放者"的区别。一个高层公司管理者对组织所有的改变总是谈论"机会"而不是"问题"。DA 著者认为人们在建构内容时都有根"柱子"作为支撑。我们每次讲话都会创建一个版本。如果问你周末做什么,那么你这次所建构的回答会与上次被问到时会有所不同,并可能与我的背景(可能是老人、男性,而不是家庭成员或者知己等)相联系。一个露天舞会可能会被描述成"有趣的"。DA 研究者有代表性地分析访谈副本,更常见的是采用著名政治家的演讲(Potter and Edwards, 1992)。

主题(或者理论导向)分析

在常规科学方法的框架中,主题分析(thematic analysis)可能以一种更加传统的方式来检验假设。通常科学必须通过某种途径以某种形式数字化才能得到认可。尽管如此,法庭经常处理定性证据(证人陈词、特征说明、精神病评定等),然而在法庭上如在科学中一样,其目的是支持这个或那个理论。我们能预言年轻的犯罪者比来自中产阶级的其他年轻人有更多的疏离感,也能用他们陈述的定性差异证明这些——其差异不是他们使用的攻击性语言的数量,而是攻击性语言类型所含的敌意。

海斯(Hayes, 1991, 1997)为理论导引主题分析(theory-led thematic analysis)提供了一个好例子,他在理论上对文化的差异作出预测,并期望在两个不同组织的工人的访谈资料中反映出这种差异。

7.4 分析定性数据——提取定量数据

我们仅仅知道定性数据可为研究提供证据，此证据在常规定量研究中用来检验假设。不过如果研究者希望从定性内容中提取定量数据，他们能够使用一种至少在1930年就已经出现的方法，即内容分析。从定性内容中提取数字数据的理由是研究结果要求精确性、可靠性和可重复性。本书前几章中列出了全部原因。这把我们带回到假设检验和更多的实证研究方法中。

内容分析的材料

可对先前存在的材料（专栏7.2中给出了一些可被分析的例子）或者研究中被试产生的材料进行内容分析。例如，我们可以要求被试完成一个故事，或者围绕一个主题如"全球变暖"写篇文章，然后提供一些有说服力的材料看是否影响他们的态度。访谈记录也能用定量方式作内容分析。

专栏7.2 心理学课题中可作内容分析的材料

材　料	分析中能识别的项目
儿童书籍	❑ 攻击性内容 ❑ 性别模式或种族刻板印象 ❑ 文化道德观点——分享、关心、帮助 ❑ 处理情感问题——死亡、爱、忌妒
儿童绘图	❑ 大小或者包括不同的家庭成员 ❑ 身体包括的部位（随年龄增长的数量） ❑ 重要物体的大小（例如圣诞节前后山特（Santa）的口袋）
电视广告或者电视连续剧	❑ 性别角色 ❑ 暴力主题 ❑ 健康饮食 ❑ 当前的道德或政治问题
报纸	❑ 不同报纸描述著名人物使用的语言 ❑ 问题的频率——谋杀、事故、抢劫
征婚广告	❑ 提及的不同性别的不同特征 ❑ 提供的内容和要求 ❑ 跨文化差异
电子邮件和因特网聊天网站	❑ 与征婚广告相关的 ❑ 幽默的体现/获得需求的途径（如办公室邮件）

内容分析研究的操作

抽 样

如何从要研究的媒体材料中取样非常关键。例如,利用《泰晤士报》研究征婚广告,同时从《太阳报》收集更多的资料并认为它们是相同的,这些做法是无用的。作为研究的一部分,你可以选择两种有差异的报纸,以比较两种不同类型报纸所做广告的差异。但必须对所选报纸进行等价抽样——用一周里相同的日子,一年中同一时期等。如果抽取电视广告样本,则需要考虑广告通常与每一档节目内容联系在一起的情况。

编 码

这是内容分析法最关键的方面,同传统研究一样,资料收集前你能决定研究所需的内容。编码通常需要参考先前的研究,考虑要检验的假设(例如,在征婚广告中,男性会更多地考虑钱的因素)和创建要使用的编码系统。达到这个目标的好方法是做一个预研究(a pilot study),抽取一些报纸,看看常规广告中都使用了哪种语言和概念。

内容分析之初需要一组严格定义的"编码单元",它们是一些通过确切途径从材料中获取的数字。有时评定者给量表的材料评分,但数据通常是频次计数(见第8章)。也就是说仅计算每一类型单元发生的总次数。在由坎伯巴奇(Cumberbatch, 1990)做的电视商业广告研究中,评定者通过计算每一特征发生的频率,得出了下面的统计资料:75%的男性和25%的女性被判断为30岁以上。男性人数比女性多2倍,89%的高谈阔论者是男性,尤其是在专门的/官方的信息方面。评定"吸引力"时,女性与男性的比率是3:2。男性忙于朋友的家务劳动可能和忙于自家一样,而女性主要忙于自家的家务劳动从来不为朋友劳动。

例如,编码单元就能被定义为:

单元	例子
词	分析不同杂志中与性别相关的词。
主题	分析儿童文学作品中的场合,或者男孩/女孩实施计划和得到赞扬的场合。
项目	寻找整个故事,如关于伊拉克的文章。
特点	分析电视中卡通形象的性格类型。
时间和空间	计算专用于媒体特殊问题的空间或时间。

步 骤

收集资料后可能要进行单元编码。在这种情况下,研究者对于创建编码单元是为了"得到一个结果"的批评是持开放态度的。尽管如此编码单元必须与被检验的假设明确相关,而没必要提供资料的非主要部分(因为它们不合适)。研究者写详细报告时要弄清楚所作的决定是什么,并且当被询问时,要准备好出示所有的原始

资料。这样就可以毫无问题地使用这种方法了。

编码系统的形成可以在资料收集前,也可以在资料收集后。要用编码系统对评定者进行训练,可以给评定者增加实践材料方面的训练直到他们之间一致(可信的),但评定者在做分析时对研究预测是不知情的。评定者之间和编码之间的信度,就像第5章中关于评分者信度的解释一样,可以通过计算同一材料中评定者间的相关获得。

7.5 个案研究

个案研究(case study)是对一个人或一个团体,例如一个公司或者组织进行深入细致的调查。被选为个案研究的人通常是由于他们的性格、能力或者经历在某些特定方面是突出的,或者是因为他们在特定人群中具有代表性,而研究者的目的是对这些人在某一重要时间段内的全面信息作一考察。这些人可能有特定的能力,如卢里亚(Luria, 1969)的撰稿人,他可以准确无误地记住长词表长达15年之久;或者他有特殊的心理条件,如西格彭和克莱克利(Thigpen and Cleckley, 1954)的"夏娃(Eve)",她在经历精神疗法时能展示三种不同的个性。他们被研究可能缘于自身独特的经历,如格雷戈里和华莱士(Gregory and Wallace, 1963)的"SB",他在52岁时经过外科手术恢复了天生的失明。最后一个例子最直接地向我们表明,对个案进行深入地定性研究,可以使一个宽泛的心理现象变得容易理解——如果那样的话,案例中的感官发展以及不寻常的机遇使得他们开始观察这个世界是什么样的。

任何一个个案研究都有其特点,它们可能在方法上有差异。如在固定不变的间隔时间进行详细的访谈非常普遍,它可能是半结构化的,也可能是非正式的。对儿童进行研究,还可能使用观察和心理量表法。运用个案研究有几条理由:

- **突出的例子**:上面的例子都属于这种类型,个案之所以被研究是因为它们是如此罕见且有自身内在的兴趣。
- **反驳一个理论**:一个独有的相反例子是对任何理论的一种挑战。如果一个失去母亲的孩子被发现在很多重要方面的发展都很正常,那么我们必须校正诸如剥夺总是毁灭性的观点。
- **数据汇总**:把从几个个案研究中收集的大量信息进行资料汇总后,可用来分析产生的特定影响。一旦这种模式形成并与其他模式相联系,那么可能实施更多的数量研究,且实施不依赖于特定的个案。
- **洞察力**:不管是否进行进一步的定量研究,个案研究在深度上的极丰富性是其独特的优势。通常我们可能想象不到个体所处的特定环境和他们应对逆境的方式。不过一旦有所发现后就能激发研究者对心理现象形成新的思考方式,也能教会我们如何更好地移情和理解,进而增加我们的心理学知识,这是不用检验特定的假设就能达到的。

练 习

1. 如果你打算进行定性研究,你将使用这里(关键术语框中所列名称)描述的哪一种方法,为什么?
2. 什么是“编码单元”?
3. 在内容分析研究项目中,如何检查评定者的信度?
4. 谈谈个案研究的优点和缺点。

答 案

1. 没有哪一种是最好的,只要检查一下它是否准确地体现了所选方法的主要规则即可。
2. “编码单元”是内容分析研究中采用的手法。例如,报纸故事中的种族刻板印象,性别或运动特征,彩色增刊中提及“全球变暖”的次数。
3. 使用统计相关(见第 12 章)。
4. 优点是所收集信息的丰富性和独特性,缺点是可能带有主观性,当研究者与非常特别和风趣的人长期相处之后,更易于产生主观的倾向。

关键术语

个案研究(case study)
诠释现象学分析(interpretive phenemenological-analysis)
编码单元(coding unit)
主题分析(thematic analysis)
内容分析(content analysis)
语义分析(discourse analysis)
扎根理论(grounded theory)

8

数据处理

——描述统计

本章内容

- 统计的一般用途，尤其是在处理表述清晰、无异议的证据（描述统计）和可能的影响效果（推断统计）方面的用途。
- 测量水平，包括称名、顺序、等距和比例测量，着重介绍与心理研究密切相关的前三种。
- 数据汇总，包括集中趋势（平均数、中数、众数）和离散趋势（全距、四分位距、标准差）。
- 图表汇总，包括条形图、直方图、频数分布表、累积频数分布表，同时讨论如何正确使用这些图表。
- 数据分布，主要是正态分布及其曲线下的面积和 Z 分数，同时简要介绍偏态分布。

8.1 统计的必要性

进行统计运算会使许多人感到头疼，但是我们必须处理所得到的数字。许多学习心理学的人认为统计非常难，感觉自己在数学方面十分无能。这可能是因为许多人将数学看成一门较为抽象的学科，认为它不像英文和地理那样与现实生活紧密相联。但是当你在超市盘算买四盒特价豆制品是否合算时，你必须运用数学。其实，本章节所涉及的运算就像买豆制品一样极其简单，你所需要的仅仅是一个能进行加、减、乘、除和开方的计算器。

首先让我们明确为什么统计是十分有用的。在心理学以及其他领域中有各种各样的数据，比如政客和广告商提供的数据以及那些需要更详尽解释的数据。面对这些数据时，如果不能对其进行正确检测和判断，那么我们就得不到任何信息。下面四个专栏是出现在媒体上的不同时期的具有迷惑性的统计数据。请试着弄清楚为什么把它们称做是迷惑性的数据（其中一些较为简单，但有一些难度较大）。

专栏 8.1　迷惑性统计 1——选择孩子的性别

据报道，一个医疗诊所出台了一项服务，宣称可以让夫妇们决定他们生育孩子的性别。在一次电台采访中，当诊所发言人被问及这项服务成功的概率有多大时，他回答说该项服务已经使六对夫妇中的四对得到了满意的结果。

专栏 8.2　格拉斯哥的诚实居民

某报纸公布了一项实验结果，实验的内容是选取几个城市，在每个城市中丢下 10 个装有现金和相关证件的钱包，最后统计出各个城市所归还钱包的个数，结果如下：

	归还	占有		归还	占有
格拉斯哥	8	2	庞蒂弗拉克特	7	3
利明顿温泉镇/华威镇	8	2	利物浦	6	4
巴塞尔顿（Basildon）	7	3	埃克塞特	5	5
伦敦（London）	7	3	卡迪夫（Cardiff）	4	6

摘自《卫报》，1996 年 6 月 17 日

记者认为格拉斯哥和华威郡明顿温泉镇的居民是英国最诚实的，因为他们都归还了 10 个钱包中的 8 个。文章作者（杰克・克罗斯利）还根据中等城市与大城市各自归还的钱包总数为 27 和 25 就得出结论：中等城市的居民比大城市的居民稍微诚实一点。你认为呢？

专栏 8.3　重大新闻！25% 的数学老师即将退休

最近一篇文章以相当惊慌的口吻写到，在未来 10 年中，会有 25% 的数学老师退休，这将导数学老师的短缺。你认为这是杞人忧天吗？

专栏 8.4　克里斯汀·迪奥的魔力凝胶

> 克里斯汀·迪奥为 Svelte 减肥膏做了一则广告，广告中宣称 550 名妇女使用减肥膏一个月，月末时有 52% 的妇女称她们的臀围减少了一寸，56% 的妇女称她们的腿围也减少了一寸。这些数据是否会促使你去抢购这种商品呢？

专栏 8.5　为什么这些数据具有迷惑性

> 专栏 8.1——如果诊所的服务是无效的，是一个大骗局，那么要在六个婴儿中正确预测四个婴儿性别的概率是多少？答案千真万确。六个婴儿中正确预测三个婴儿性别的概率是最大的。对于每个婴儿，诊所都有 50% 的概率使其符合夫妇们所选择的性别，也就是说在六对夫妇中，最有可能有三对夫妇得到所选择性别的婴儿，而其他三对则失败了。事实上，这个概率为三分之一，假如有四对夫妇成功的话，则其概率为四分之一 ——这种概率很小因而不太显著。
>
> 专栏 8.2——与预测婴儿性别一样，这里也存在着样本大小的问题。10 个钱包归还 4 个与归还 8 个的确有很大差别吗？代表整个格拉斯哥或卡迪夫的样本仅仅是 10 人吗？假如卡迪夫再有 4 个人归还钱包，卡迪夫的居民就与格拉斯哥的居民一样诚实吗？从一个小样本推及到总体的跨度太大了，而且假定所有的未归还者不诚实也太绝对了，他们可能是太忙或者仅仅是忘了。
>
> 专栏 8.3——任何职业的工作时间跨度都约为 40 年，假定人口均匀地分布在各个年龄群体中，那么任何职业都将约有 25% 的工作者处于 50 岁至退休年龄之间。
>
> 专栏 8.4——这里的问题确实有点复杂。以广告形式呈现的这些数据看似十分具有说服力。但是我们并不知道迪奥怎样得到这些数据的/他们问那些妇女的问题是什么。我猜测他们是让妇女们在月初测量一次，月末再重复一次。但这样的测量将会怎样呢？其实，当对某物测量两次时，其测量结果之间必定会出现细微的差别。之所以出现差别，原因是机体的状态是不断变化的，而且每次测量的方式不可能完全相同。即便你测量的是木头或钢铁，两次测量结果之间也会存在细微的随机差别。正是因为测量中存在细微的随机差别，我们应该明白 50% 的妇女的确瘦了一点，但还有 50% 的妇女即使不用减肥膏也会瘦一点。此外不要忘记，当第一个被试群体没有得到理想的结果时，商家可能会选取另一个 550 人的群体。商家们只是将广告研究作为商业秘密却不重视科学研究的客观原则。当然，并不是说商家们一定是这样做的，只是说我们都不太清楚他们究竟是如何做的。

描述统计与推断统计

统计有两个既有联系又有区别的目的。首先，统计用来描述发生了什么。正如当问你在过去几年中高速公路上发生了多少起事故时，你会将各年的数据逐一展示。上面提及的钱包归还的研究中，统计数据告诉我们各个城市钱包归还的数目。像这样的数据就叫描述统计，本章的任务就是学习怎样处理各种数据资料以对其进行正确清晰的描述。

先前所述的各专栏还表明，样本还含有其总体的各种信息。记者根据 10 个人

的样本来推断各个城市的居民在诚实度上有差异；声称能决定婴儿性别的诊所以及通过数据证明纤体凝胶能减肥的事例也都是由样本推论的。当我们由样本推论总体状况时就要使用推断统计，这些是 9～12 章的内容，届时我们将学习如何通过一个相关的小样本对总体进行推论。推断统计是用来检验假设的，例如检验纤体凝胶能减肥。但心理学上的假设可能是听音乐时会增加词语加工任务中的拼写错误。

结果与结论

描述统计是描述研究中出现的事实，是研究的结果。但是我们感兴趣的是对结果进行概括并给予结论。区别结果与结论这两个术语是很重要的，因为在与考试有关的问题中经常涉及它们。如果被问及研究者可以从研究中得出什么结论时，你应该仔细考虑而不要简单地重复结果。例如，研究者选取一个 20 人的样本来完成一项复杂任务，被试在高温环境中会比在低温环境中表现出更多的言语暴力和更少的笑容，那么该研究的结果就是描述从 20 个被试身上获得的有差异的数据。结论可能就是高温会增加人的攻击性（一般情况下）。

8.2 出发——测量水平

首先，必须确定收集到的数据类型。所有收集到的相关资料都具有一定的测量水平，它决定了在检验假设时使用的推断统计方法。学习之前先看下面几个例子。

1. 吉纳个子矮。	称名数据
2. 吉纳比简矮，但比萨曼莎高。	顺序数据
3. 吉纳的身高为 167 厘米。	等距/比例数据

第一个例子中，吉纳属于矮个子的一类人，这是与其他类型相比较的，比如说还有高个子和中等个子。第二个例子中，我们得知吉纳、简和萨曼莎相比处于中间位置，但只能知道他们三个人的身高排序，却不知道吉纳比萨曼莎高多少。第三个例子中，能确切得知吉纳的身高，假如使用统一标准的尺度进行测量的话，便可将其身高与任何人比较。我们曾在第 6 章提到，用统一测量尺度来进行心理测量是许多心理学家的梦想。下面就对上述例子中右边所列的数据名称进行解释。

称名测量——分类变量

处理不同质的数值时要使用称名测量。实际上这里并没有真正的测量而只是简单地统计一下各个类别的数量，也就是我们所熟悉的频数。假如要对儿童卧室的玩具进行研究的话，我们不会去测量他们玩具车的长度，合理的研究方法是统计一下他们拥有的玩具车、棋盘游戏、电脑游戏、运动用品、玩具刀枪、玩具娃娃以及其他玩具的数目。表 8.1 就是典型的称名或分类数据资料。

表 8.1 大学食堂顾客统计

顾客类别	人 数
学生	650
教学员工	34
非教学员工	45
外来人员	12
其他	3

要确定数据资料是否是称名数据，你可以问自己一个简单的问题：

我从每个受测者身上得到了什么信息？

如果你得到的信息类似于“他们是 X、Y 或 Z”，那么你得到的是称名数据。同样，如果你只能简单地将某人与其他人划分到某一类别中却不能以某种尺度进行区分，那么你得到的还是称名数据。

例如对学生的研究经常涉及某些观测变量在性别上的差异，比如说拿书时是将书放在体前还是体侧。还有一个常见的研究变量是汽车司机进入停车场时是倒着开进去还是正着开进去。上述例子中性别变量是分类别的，你不是男性就是女性（除去某些性别偏差和生理异常）。同样第二个变量也是分类别的，因为你开车进停车场不是正着开进去就是倒着开进去，研究的结果如表 8.2 所示，这些数据资料称为称名数据。

表 8.2 不同性别司机进入停车场方式的统计（称名数据）

	男 性	女 性
正开进入	85	72
倒开进入	45	43

顺序测量

如果老师让你们班的 20 名学生按身高分成两个等组，一个高个组一个矮个组。你们可能在两组中不断权衡调整直至最后分成两组。但是，由于自尊心强弱的影响，有的学生对于自己处于矮个组感到不满，因为他们矮个组中是最高的，所以他们并不认为自己矮。之所以会出现这种问题是由于（除了他们的自我）使用的类型变量仅有两个类别，正如上文中吉纳的例子。现在假如老师让你们按身高排列的话，我们将得到更多的信息。

如果数据资料以顺序水平整理的话，那么群体里每个成员的位置就可以确定了。上文中自尊心脆弱的男孩可能对自己身高排第十感到好受一些，然而我们依然不知道他比更矮的第九个学生高多少[①]，因为顺序数据并没有说明等级之间的具体差距是多少。例如在一场自行车比赛中，你只落后第一名 0.1 秒却只能得第二名，你们领先其他选手 10 千米，结果，下一个人落后你 10 分钟却得到了第三，这确实有点令人不解。

① 应该是排第十一，不是第十，英文版错误。——译者注

如何对数据进行等级排序

通过问卷得到的分数可能像表8.3那样，并不是顺序数据（虽然有些教师或教材称其为顺序数据）。顺序数据总是像表的第三列那样以等级形式呈现的，但是数据资料通常并不是顺序数据也无需必须转化为顺序数据，这完全是由我们自己决定的。下面将讨论什么时候才可以将数据作为顺序数据。

表8.3 等级变化分数

姓 名	分 数	等 级	高(H)或低(L)
安	18	5.5	H
贝思	25	7	H
卡罗尔	14	1	L
唐	18	5.5	H
埃玛	15	3	L
弗恩	15	3	L
吉尔	15	3	L
海蒂	29	8	H

分析中我们没有将等级1作为最高分而将其设为最低值。因此，表8.3中的14分对应于等级1。埃玛、弗恩和吉尔共同占据下面的三个位置——数学上可以称之为“并列第二”，但统计中我们只简单地将它们按照其占据的三个等级的位置进行分配，即共享第二、第三和第四的位置，因此他们的平均等级就是 $(2+3+4)\div 3=9\div 3=3$。简单的法则是：如果共享的等级个数为奇数（就像这里）就取中间等级。如果共享等级个数为偶数那么就取中间两位置的中值。我们可以将这个规律应用于安和唐[①]这两个等级为5和6、得分同为18分的人身上。这两数的平均值为5.5，所以5.5就是他们的等级。最后，还有等级7和等级8，分别对应于25和29分。如果有四个人共享等级6、7、8、9，那么他们每个人的等级都是7.5。

等距与等比测量

顺序测量的缺点是从数据中我们无法得知一个人的得分和另一个人相差多大。但在**等距**测量中我们可以得到这些信息。等距量表使用相等的单元，量表上10分与20分之间同40分与50分之间的数量是一样的。这样无论是采用英尺或是米做单位，在时间和空间距离上都是准确的。重量是另一个例子。

心理测量专家的想法（第6章）是心理量表可以产生这类数据。如果这种情况是正确的，那么IQ得分为100的简，就领先于IQ为90的彼得，就如IQ为125的迈克尔领先于IQ为115的休(Sue)。当然，我们无法将智力等距化。在心理学中，实验的实施者所能采取的最好的方式是：如果它们是真正的等距水平，就可以制作量表模拟将要发生的事。如果测量是等距的，则可以形成正态分布，就像身高一样。如果心理测量能形成这样的分布，那么这些数据就可以作为等距测量数据。图8.1描述了不同测量水平之间的比较和每种测量水平所提供的信息。注意有些等距量

① 从表8.3中可看出应是安和唐，而不是弗恩，英文版错误。——译者注

表是离散的，它们在整个数据中的任意两者之间都没有值，就如同不能拥有 2.4 个孩子一样。你可以有 2 个和 3 个孩子，但无论哪个父母都不会说自己的孩子有半个脑袋。

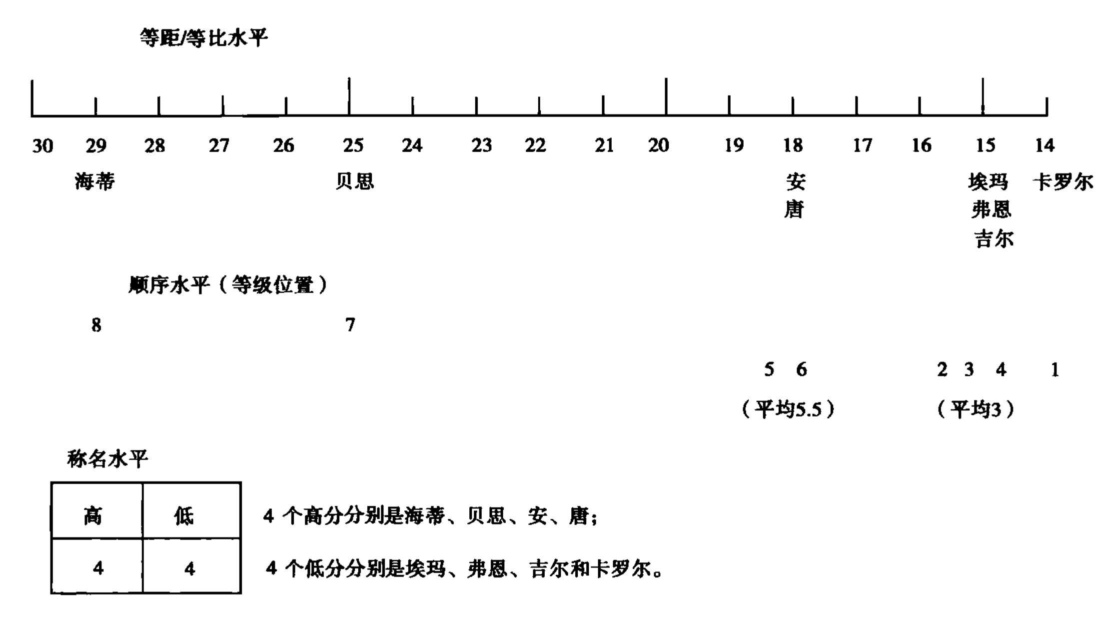


图 8.1 测量水平与所提供的信息（基于表 8.3 的数据）

课程大纲或许要求理解比例测量这一术语。等比数据存在于具有绝对零点的等比量表上，量表上的比例是有意义的。举例说，温度的测量，并非开始于真正的零点。30 ℃不是 15 ℃的两倍那么热；数值虽是两倍但热值并不是。如果将其转换为华氏温度就很容易看出，它们分别对应于华氏 86°和华氏 59°。温度的测量是等距测量但不是等比测量。同样，心理测量大多数也是等距测量而非等比测量。然而，心理学家从不担心他们的数据是否是等比水平，他们在分析时并不将等距数据和等比数据加以区分。对绝大多数分析过程来说，等距数据已经足够。因此，从现在开始，我们将不再提及等比数据，而是将所有的数据看做等距数据并且假定平均等距或等比。

何时将"等距"数据作为顺序数据对待

心理学中，断言量表分数是真正的等距分数通常是不合理的——即使是标准化的态度量表或智力测验，尽管心理测量学家们希望如此。我们通常会采用如同等距量表那样表现良好的评分系统，这种系统上被称之为准等距量表。只要数据大致服从正态分布，对于诸如"用量表上的 1 到 20 这些数值显示你自己有多虔诚"之类的评分也是真实可信的。然而，一般来说，如果一种量表被定义成非标准化的，就像刚才提及的那种，或是简单地让人们在任意情况下进行数字判断，通常最可靠的办法就是通过等级划分，将评分结果作为顺序数据加以处理。例如，当你请人在一个从 1 到 10 的量表上对画中人的"吸引力"进行评估的时候，就需要用到这种方式。

数据转换及信息缺失

数据在它们第一次被测量的水平上并不是固定的。我们可以将数据从等距量

表向下转换为另一种量表,但是向上则不行。这就是为什么在提出假设之后,在收集数据之前,就要非常仔细地考虑该使用何种分析手段的原因。你需要确认测量工具可以收集的数据的水平。有一次,我的一个学生做了一个实验,让所有被试阅读关于攻击一位妇人受到攻击的故事,其中一半被试阅读的故事中这位妇人"穿着朴素";另外一半被试阅读的故事中妇人则"衣不蔽体"。被试需要回答的问题是"谁应该为攻击受更多的责备:攻击者还是妇人?"当然,所有的参与者选择了攻击者。可以看出,由于这位学生确定的选择类别只有两类而且没有给被试熟悉的空间,因此她所采用的测量系统就显得太过粗糙。她至少需要采用类似"从0到10的量表"对攻击者的受责备程度进行评判,同时对妇人也采用这种方法。这样的话,她就可以看到是否在"衣不蔽体"这种情况下被试对受害者的同情有所减少。后来这个学生又邀请所有的120位被试重新来做此项实验。

将数据向下转换通常是可行的,但这种方式会丢失信息。表8.3显示了数据组从等距向顺序水平的转换。右面的最后一列是转换为称名数据,这里我们将所有的分数从中间分开,一组称为"高分组",另一组称为"低分组"。当然,我们并没有对每种分类中的所有参与者加以分离。我们对顺序数据进行了分离但却并不知晓两个等级位置之间的真正距离。

练　习

1. 一组外科诊断记录上将病人分为"慢性的""急性的"和"未分类的"。采用的是什么测量水平?
2. 对滑冰比赛中的编排分和表演分进行评判时,什么样的测量水平对数据来说是最可靠的?
3. 看一下表8.4, a,b,c和d列中都有测量。这组数据分别采用了何种测量水平?
4. 表8.4中从a到d,哪一列包含了测量水平中的大多数信息?
5. 你的姐姐声称,因为她今年在班级里的三次数学测试中都得了第一,她肯定比其他学生优秀得多。她的话里有什么错误?(你敢指出吗?)
6. 用三种办法测量驱动能力,一种使用称名数据,一种用顺序数据,一种用等距数据。
7. 你能将表8.5中的数据先转换为顺序数据,再转换为称名数据吗?你需要填写一个空白表格。提示:对于顺序数据,将所有的值看做一组。对于称名数据,试着将其按高/低进行分类。
8. 下面是测量因变量的几种方法。每种方法都决定了其所使用的测量水平。选择:
   (1)称名型;(2)顺序型;(3)等距型
   (a)人们被问及在毕加索(Picasso)、马蒂斯(Matisse)或戴利(Dali)三个人中,喜欢谁的画?
   (b)压力问卷中各种各样的职业规范是如何建立的?
   (c)让参与者根据吸引力水平即从"最吸引"到"最不吸引"对照片排序。
   (d)参与者对各种线的长度的估计。
   (e)将卡片按时间进行分类。
   (f)人们的选择:《太阳报》《泰晤士报》还是《卫报》。
   (g)参与者根据从1到10的量表评估自我价值和自我估计。
   (h)参与者在卡特尔的16PF量表中的得分。
   (i)通过照片测量并区分与两个参与者进行私下交谈时他们的立场。
   (j)用1~10评判每个参与者对其生命中关键事件的理解程度。

答 案

1. 称名型。
2. 顺序型。(人为评定)
3. a. 顺序型。
   b. 等距型。
   c. 称名型。(或种类)
   d. 准等距型——称顺序型更可靠。
4. b 列。
5. "最高"是顺序量表。我们并不知道她比别人要优秀多少。
6. 例:称名型——是否撞到围栏;顺序型——通过平稳度训练后的位置;等距型——测量比赛中的速度。
7. 顺序水平

一致性	不一致性	称名水平	一致性	不一致性
6.5	11	最快 9	6	3
9	17	最慢 9	3	6
1	11			
2.5	4.5			
16	18			
2.5	4.5			
6.5	8			
11	14			
13	15			

8. (a)称名型
   (b)等距型(由于标准化)
   (c)顺序型
   (d)准等距型
   (e)等距型
   (f)称名型
   (g)准等距型
   (h)等距型(由于标准化)
   (i)等距型
   (j)顺序型

关键术语

推论(conclusion)
描述统计(descriptive statistics)
离散数据(discrete data)
结果(findings)
频数(frequency)
推断统计(inferential statistics)
等距水平(interval Ievel)
测量水平(leve of measurement)
称名水平(nominal level)
顺序水平(wrdinal level)
准等距量表(quasi-interval scale)
等比水平(ratio level)

表 8.4 英超表(顶级球队)

英超球队	a 排名	b 积分	c 地区1 = 伦敦 2 = 北方 3 = 南方	d 支持率 (虚构)
切尔西	1	95	1	1
阿森纳	2	83	1	3
曼联	3	77	2	6
埃弗顿	4	61	2	4
利物浦	5	58	2	2
博尔顿	6	58	2	5
米德尔斯堡	7	55	3	8
曼城	8	52	1	7

表 8.5 巩固练习 8.1 中练习 7 的阅读时间

a)阅读时间/秒		b)顺序水平		c)称名水平
连续故事	非连续故事	连续故事	非连续故事	
127	138			
136	154			
104	138			
111	117			
152	167			
111	117			
127	135			
138	149			
145	151			
平均时间:$\bar{x}$ = 134.3				

如何总结数据

描述统计的重点是找到一种有效、清晰,并且公正的数据表达方式。我们不需要呈现获得的全部原始数据。原始数据是我们从每个参与者中得到的未处理的数字,假定我们已经对 50 个人在两种不同照明条件下的反应时间进行了测量,将 100 个未处理的分数公布给读者无论如何都是没有用的。我们需要为读者总结数据以便于给出任何反映某种趋势的、清晰的图形或是发现的不同之处。例如,在这种情形下,我们也许想要看到在每种条件下的平均得分。

我却不会作出总结

本书的许多思想都是源于你以前从未怀疑过的日常普通概念。即使你憎恨数学、畏惧统计并且从未在这个领域内做过任何正式工作,在你的生活中肯定会有很多无意识的统计描述。你可能相信只有聪明的、有数学头脑的人才能做这类事情,但请考虑这一点:假设你刚刚在一所新的大学上完第一节课并在餐厅和我们相遇。

如果让你对其他同学加以描述，而你决定从他们的年龄开始。你不会一开始就告诉我每个班级成员的确切年龄，这可能会花很长时间。你很可能会说一些诸如"班上大多数同学在25岁左右，但有两个十几岁的或是有一两个超过40岁的"。实际上，你已经从统计的角度概括了班级成员的年龄，只不过不太精确而已。首先你给出了一个大致的平均数以及班级里的典型年龄，然后你让我了解了当前班级里典型年龄的变化。这两个概念是对测量数据进行统计描述的基础。这是一个从概念上发展而来的例子，它存在于日常生活中，并出于统计目的对同一件事情更形式化的说法。下面有两个一般性的正规术语用于描述成员年龄的特点。

集中趋势：它不同于"中间"的含义，而是指一组数据中最集中的或最典型的数值，比如上例中的25这个数值。一般来说，集中趋势没有平均数严谨。但在描述统计中，我们需要明确平均数的某种类型。

离差：用来计量一组数据中所有数据偏离中间值或典型值的程度。这是一个重要的概念，如表8.6所示。7人来自有着猎狐历史的农村社区Dunton Parva，7人来自大城市的贫穷地区Slumditch，对于狩猎的范围他们已有的态度值为40。每个组的平均值是相同的，但要看每个组的变化。看上去Dunton Parva社区的人在数值上有所分化，而Slumditch社区的人并没有什么差别——他们将会怎样做；你不可能在Slumditch社区对狐狸进行大量的狩猎！图8.2表明两种变量或离差的相关大小。

表8.6 DP和S社区对狐狸狩猎的态度得分情况

Dunton Parva 社区	Slumditch 社区
38	19
12	27
36	23
8	24
25	21
34	22
9	26
23.1	23.1

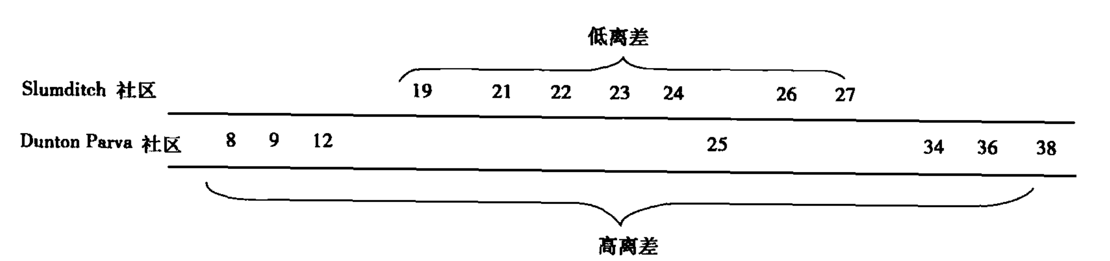


图8.2 狐狸狩猎分数的不同离差范围

8.3 集中趋势测量

平均数

平均数就是我们平常所说的平均数值。为了得出一年之内每个月的电话费用，你要拿出 12 个月的所有账单，把它们相加然后再除以 12。

专栏 8.6 五种变位词解答时间的平均数

17
10
23
12

总和 = $\frac{13}{75}$ 秒

平均数是所有数值的总和除以数值的数目。因而如果把五个人的数据 17,10,23,12 和 13 相加，我们会得出总数为 75，并且有 5 等份，因此平均数是：

$\frac{75}{5}$ = 15 秒

平均数公式

让我们切入正题——在绝大多数统计中都需要使用公式。公式就是一套简单的经过编码的说明书——就像食谱——告诉你做什么能得出某个统计数字。尽管每次统计运算都需要公式，但并不用为此担心，因为大量的统计程序中公式都是存在的。



图 8.3 使用公式如同使用菜谱

平均数的公式是：

$$\bar{x} = \frac{\sum x}{N}$$

$\bar{x}$通常用来表示平均数。$\sum$ 是希腊字母 S，用来表示所有数值的求和。x 用来指代每一个数值，$\sum x$ 通常表示所有数值相加。N 是一组数据中数值的个数。本章最后有类似的小结。表述结果时通常用 M 表示平均数。

平均数计算步骤

1. 将所有数据相加——即 $\sum x$。

2. $\sum x$ 除以该组数据的个数 N—— 因此得到$\frac{\sum x}{N}$ 。

平均数的优点

- 平均数是用来估计总体参数的重要统计数据，它是进行显著性差异检验或相关检验等参数检验的基础。
- 平均数是三种集中趋势测量方法中最敏感和准确的。（因为它是等距测量，并且考虑到一组数据中数值间的确切距离）

平均数的缺点

- 由于平均数很敏感，因而容易受个别不明数据的影响。例如，6 个人进行变位词测试，第 6 个人的得分为 225，那么平均数将变为：

$$x = \frac{17 + 10 + 23 + 12 + 13 + 225}{6} = \frac{300}{6} = 50 \text{ 秒}$$

225 不是这组数据样本的代表值，其他 5 个人解决变位词的速度都要比它快得多。那些处于一般范围之外的数据就是极端数值，就像这组数据中的 225，极端数值会严重影响平均数，见图 8.4。

平均数还有一个小缺点，就是当把从离散变量中得到的“愚蠢的”数值作为平均数时，它会误导，至少会扰乱人们。例如那个臭名昭著的例子即一对父母有 2.4 个孩子。

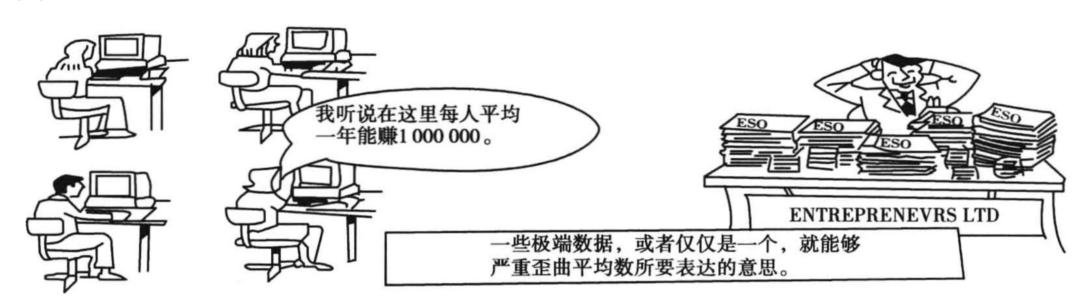


图 8.4 极端数值会使平均数失真

> 进一步运算前需要注意小数和伪精确性,请不要使用5.428 57!
>
> 学生的答案中经常会出现(有时候是因为使用了不准确的计算机程序)9,8,7,5,4,3 和 2 的平均数为5.428 57。但后面的四位小数并没有增加精确性,原因在于原始数据是整数而平均数用小数就会不精确。通常的做法是保留小数点后一位小数。因而这里的平均数应为5.4。

中位数

与平均数不同,**中位数**不受极端数据的影响,它是一组数据中的中间数值。

中位数计算步骤

1. 把 N 个数据按大小顺序排列:15,7,12,20,18→7,12,15,18,20
2. 找出位于中间位置的数值 = $\frac{N+1}{2}$。上面例子的中位数为 $\frac{5+1}{2}=3$。
3. 如果 N 是奇数,处于中间位置的数就是这组数据中的中位数。在上面的例子中,中间位置是3,因此从左边起第三个数据就是中位数,即15。它是一组数据的中间值。
4. 如果 N 是偶数,中间位置处于两个数值之间。例如有六个数值,中间位置就是 $\frac{1+6}{2}=3.5$。3.5 这一数值表明,中位数就是该组数据中第三个数值与第四个数值的平均数。例如3,6,9,12,20, 25 这组数据的中间位置是 3.5,中位数就是第三个数9 和第四个数 12 的平均数,即 10.5。

如果中位数处于一个系列数值中,比如2,3,5,5,5,5,5,8,8,9 数据组中的中位数就位于5 个5 之间,这种情况有更精确的方法计算中位数,不过这要取决于中位数处于数据组的位置有多远。但实际中上例中的5 已经是一个相当可靠的中位数。

中位数的优点

- 不受极端数据和模糊数据的影响。因而更适合应用于偏态分布——见本章最后。
- 比平均数容易计算(提供互不联系的小组,或者是忽略那些数值相联的情形)。
- 极端数值未知时也可以计算。例如,给一组数据增加两个极大值,中位数很少受到影响。

中位数的缺点

- 无法估算数值间的精确距离。
- 不能估计总体参数,见本章后面部分。
- 在小数据组中不具有代表性,比如2,3,5,98,112 数据组中的中位数是5。

众 数

众数就是一组数据中出现次数最多的那个数值。比如在一组数据 1,2,2,4,5,6,6,6,7,8,8,9 中,众数就是6,因为6 比其他数值出现的次数多。2,2,3,5,6,6,7

这组数据被称为双模型，因为它有两个众数。众数就是出现频率最高的数值。一个容易犯的错误是把数据组中数据出现的次数当作众数。例如，表 8.1 中众数是“学生”而不是这组数据中出现的学生数目 650。

众数的优点

- 能表明数据资料中经常出现的或典型的数据。
- 不受某一方向的极端数据的影响。
- 当极端数据未知时也可以得知。
- 当数据分布范围分散时比平均数能提供更多的信息。

众数的缺点

- 无法估算数值间的精确距离。
- 不能估计总体参数，见本章后面部分。
- 对某些数值按相同频率出现的数据组来说，众数没有意义（1，1，2，3，4，4）。
- 对于双态分布来说，需要两个众数。
- 当数据被组合在一起成为如同图 8.11 中的组距时不能精确估计。但若仅仅是要求选择一个区间时，我们可以得到一个频数区间。

集中趋势测量及测量水平

- **等距/等比测量**：平均数是最敏感的，但至少要适用于等距测量，否则根据量表分数计算平均数就无法等距，结果将不准确。
- **顺序测量**：如果数据不等距但可以排序，表示集中趋势则可以用中位数。
- **称名测量**：如果数据来自于不同的类别，那么只能用众数。
  众数也可用于顺序和等距数据。
  中位数也可用于等距数据。

8.4 离中趋势测量

图 8.6 显示，Dunton Parva 地区居民的态度得分比 Slumditch 地区的居民有更多的变化。变化这一概念是所有数据的特征并且在统计分析中起着基础性作用。尽管工厂中的设备都是根据程序制造相同尺寸的产品，但它们也有细微的变化。不同的黄瓜有不同的长度。人身上通常在身高和体重方面有一些明显的变化，但这种变化更多地体现在一些心理变量上，比如技能、态度、人格特征等。因而，我们需要对组内变化的数值加以测量。

全　距

全距就是一组数据的最大值和最小值的距离。

全距计算步骤

1. 找出一组数据的最大值和最小值。在专栏 8.6 中分别是 23 和 10。

2. 计算最大值和最小值的距离。在专栏 8.6 中是 13。

3. 把第二步的结果加上 1。在专栏 8.6 中是 14。

为何加 1 呢？当测定人们要花多久去猜字谜时，我们无法说他们正好用了 23 秒。对于所有像这样的等距测量我们只能知道他们大致花了 22.5 到 23.5 秒。最小值也可能是 9.5，所以全距就是 23.5 - 9.5 = 14。当我们用等距量表进行测量时，所测量的是最接近的区间而不是精确的点值。这个方法即使当量表不是完全等距时也适用，这主要是由全距这个概念本身决定的。

从表 8.6 中的数据我们可知道，Dunton Parva 得分的全距是 31 而 Slumditch 的全距是 9。他们的集中趋势相同但离中趋势却相差很大。

全距的优点

- 包含极端值。
- 容易计算。

全距的缺点

- 任一端的极端值都会引起歪曲并因此产生误导。
- 不能代表两极端值之间的数据分布的任何特征。例如，全距不能告诉我们数据是掺杂着一些异常值而紧密围绕着平均数，还是总体上遍及整个范围的。

四分位距

离中趋势测量中比全距更加灵敏的是**四分位距**。我们只要根据数据的大小排序（不要忘了这个要求）计算出该序列四分之一的点值（第一个**四分位数**）以及四分之三的点值（第三个四分位数）就可以了。记住我们之前计算的处于序列数据中间位置的点值称之为中位数。四分位距就是上面两个四分位数之间的距离。不论你是否相信，半四分位距就是四分位距的一半！请看下面一组数据：

7 7 9 12 14 16 16 17 19 20 22

9 是第一个四分位数，19 是第三个四分位数，因此四分位距是 10（且半四分位距是 5）。

四分位距的优点

- 告诉我们中间数据如何反映平均数。
- 不受任何一端极端数值的影响。

四分位距的缺点

- 没有考虑数据组的高端值和低端值。
- 不能用于参数检验（见第 11 章）。

离差分析

从表 8.6 可以看出，Slumditch 中的所有样本数据都紧密地围着平均数 23.1。Dunton Parva 中的样本尽管有同样的平均数但数据却很分散地围着平均数。因而

在一组数据内存在差异。现在来看图 8.5 中描述的分数，如果想比较小组中有多少人偏离了平均数，我们就需要用**离差分数**(deviation score)。离差分数简单地说就是其原始分数偏离平均数有多远。那么，如何求离差分数呢？先根据原始分数计算出平均数。你应当采取这种方法而不是其他方法，否则会导致错误的结果。计算离差分数的公式是：

$$d = (x - \bar{x})$$

其中 x 是原始分数，$\bar{x}$是平均数。

如果用 Dunton Parva 的最高分 38 减去平均数，就会得到 $d = 38 - 23.1 = 14.9$，这一离差范围如图 8.5 所示。这是最大的离差，比 Slumditch 中的任何离差都要大。它是一个正值，因此落在图表中平均数的右侧。Slumditch 中的最小分数是 19，用同样的方法(记住根据得分得出平均数)可以得到离差分数为 $19 - 23.1 = -4.1$。这一数值是负的，因为它小于平均数，因而落在图表中平均数的左侧。

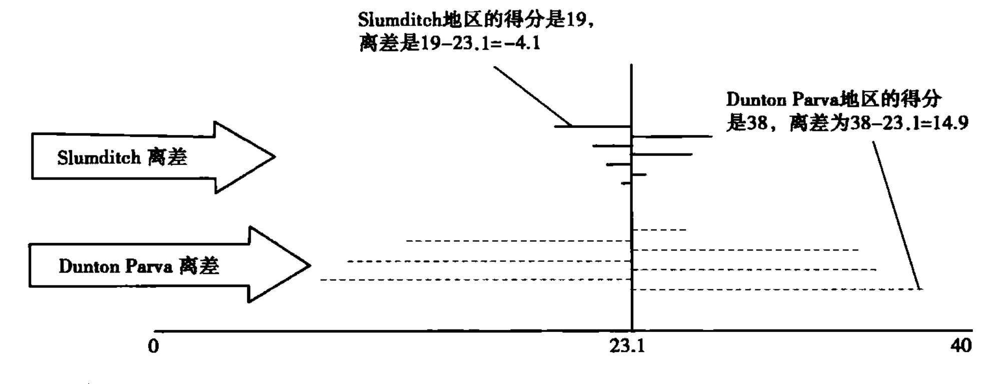


图 8.5 Slumditch 与 Dunton Parva 态度得分的离差比较

如果想指出一组数据中的数值变化有多大，只能用所有离差分数的平均数。我们需要像处理正值那样来处理负值，否则所有的数值将互相抵消。我们只是运用了真实的离差大小，而不是考虑它究竟落在平均数的哪一侧。Dunton Parva 的平均离差可能很高而 Slumditch 则相对较小。一个数据组中所有绝对离差的平均数就是所谓的**平均差**(mean deviation)。平均差在心理研究中很少使用，不过库里坎(Coolican,2004)给出了计算平均差的步骤。平均差是把所有的离差值都作为正值加以处理。

标准差与方差

由于某些奇怪的原因，统计学家更倾向于用**标准差**(standard deviation)进行组内的离差测量。也许仅仅是因为他们喜欢那些看似深刻的计算。如果没有计算器的话，很难求出平均离差，标准差也同样如此。

事实上，标准差对统计学家来说有着更多的意义。有一个很好的理由可以说明为什么社会科学中通常用标准差进行统计。我们会在本书的后面探讨参数检验及其在正态分布上的可靠性等话题。标准差则与此紧密联系，并且对从样本数值中估计总体数值非常有用。

希望你有一个计算器并能计算出标准差，或者可以利用计算机进行计算。然而，如果你不得不用手工加以计算，此处会告诉你如何去做。首先来看计算标准差的公式：

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}$$

注解：当 $d=(x-\bar{x})$（解释如上）时，$s=\sqrt{\frac{\sum d^2}{N-1}}$ 看起来更容易一些。

注解：**方差**（variance）就是标准差的简单平方，因此在上面的等式中没有平方根符号。

标准差计算步骤

上面的小公式能告诉我们该如何去求标准差吗？记住，公式不过是一套说明书。它告诉我们先求出每一个 d（与 $x-\bar{x}$ 相同），再对 d 加以平方（自己乘以自己），然后把所有的 d **平方**相加，用 $N-1$（样本数减 1）除相加的结果，最后开平方根。开平方根与平方是相反的过程，任何计算器都能够运行（例如 9 的平方根是 3，是因为 $3\times3=9$）。让我们根据一个实例一步一步地计算标准差。以表 8.6 中 Dunton Parva 的得分为例：

步　骤	表 8.7 中的数据计算
1. 求出数据组的平均数	$\bar{x}=23.1$　见表 8.6
2. 用数据组中的每一个数值减去平均数，计算每一个的离差数值（$x-\bar{x}$），用 d 表示	见表 8.7 第 2 列
3. 把每一个 d 都进行平方	见表 8.7，第 3 列
4. 求离差平方的和（$=\sum d^2$）	$\sum d^2=1\ 060.87$（表 8.7 第 3 列的总和）
5. 用 $N-1$ 除第四步的结果	$\frac{1\ 060.87}{6}=s^2=176.81$（$s^2$ **是方差**）
6. 求第五步的平方根	$\sqrt{176.81}=s=13.30$

Dunton Parva 得分的标准差是 13.30[①]。

如果你运用计算器，要确信使用的标准差公式版本在等式底部有 $N-1$。大多数统计计算器会显示像 σ_{n-1} 这样的符号。

标准差公式中的符号是：

s 用于计算样本标准差。

σ 用于计算整组或总体标准差。

进行结果报告时使用“*SD*”。

离差分析优点

- ❑ 告诉我们所有数据离开平均数的情况。
- ❑ 可用于参数估计和检测（见 11 章）。

① 必须在最后一位数字后加上 0 以表明进行了十进位小数运算——计算结果是 13.296。

离差分析缺点

- ❑ 易受数据组中任何一端的极端分数影响。
- ❑ 比其他离差测量更难以计算。

表 8.7 Dunton Parva 态度得分的部分标准差计算

得分(x)	$d=(x-23.1)$	d^2
38	14.9	222.01
12	-11.1	123.21
36	12.9	166.41
8	-15.1	228.01
25	1.9	3.61
34	10.9	118.81
9	-14.1	198.81
		$\sum d^2 = 1\,060.87$

总体参数与样本统计量

选举期间,你会经常看见投票意向调查宣称选举结果精确到"+或-3%以内"。这些调查由那些能在获得选举权的结果后得到丰厚利益的商业组织实施。倘若结果是错误的,新闻报纸就不会付报酬给他们。他们采用一个小到1 500人的样本加以调查,并根据调查结果精确预测所有投票人此刻的投票意向。心理学中,我们通常并不关心为了一个具体的研究而获得的样本特征。像民意测验专家一样,我们感兴趣的是样本所代表的总体特征。假设一个研究者推论宠物可以使人的心情平静下来,因此那些有宠物的学生在考试中的表现要好于没有宠物的学生。不可能对全部的有宠物或无宠物的学生进行测试。我们需要做的是,根据有宠物和无宠物这两个指标选取等量的学生样本并测验他们的考试成绩。如果结果像假定的那样二者之间有相当大的差异,这样就可以推论宠物对总体学生的考试成绩有影响。

适用于样本的统计当然称之为统计量,但适用于总体的统计则称为**参数**(parameters)。推论出总体参数大小的测试称为参数估计,见11章和12章。在上述计算标准差的公式中,运用了$N-1$这一数值。如果求样本平方差则需要用N,如同求任何数据的平均数一样。但是,参数估计中运用的标准差是对总体标准差的估计。进行估计时,我们喜欢通过较大的估计以允许少量的容忍度(即统计学家所称的抽样误差);这需要通过使式子下部的数值更小($N-1$,而不是N)来实现。当选取一个样本并从中估计总体平均数或标准差时会出现**抽样误差**(sampling error)。我们总会或多或少出现这样的"错误"。



图 8.6 宠物能提高学生的考试成绩么

练 习

1. 下面记录的是进行猜谜任务所用的时间(以秒记)

   12,8,23,13,17,15,18,21,18,14,18,29,55,12

   ①指出用哪个集中趋势量更为恰当?

   ②计算平均数、中位数和众数。

   ③为什么平均数要略高于中位数?

   ④计算全距、四分位距和标准差(只有你想计算标准差,并且它出现在你的教学大纲里时才做计算)。

2. 有一组数据包括 3.2 这一数值,标准差为 0。你能否猜出平均数以及其余的数字各是多少?

3. 一个政治家愤怒地声明"多达一半的孩子在平均阅读年龄以下"中有点奇怪的是什么?

答 案

1. ①该测量是等距测量,因此用平均数是恰当的,但你或许注意到平均数会受到极端分数 55 的影响,因此用中位数也许更好一些。

   ②平均数 = 19.5;中位数 = 17.5;众数 = 18。

   ③如上所述,平均数易受极端分数 55 的影响,而中位数则不会。

   ④全距 = 48;四分位距 = 21 - 13 = 8;标准差 = 11.49。

2. 如果没有任何离差,那么所有的得分都应该是一样的。因此所有的得分都为 3.2,平均数也为3.2。

3. 如果没有出现大量偏态,我们总是期望任何一组中的一半得分低于组均值和一半得分高于组均值。它是一种集中趋势的度量。

关键术语

集中趋势(central tendency)	参数(parameter)
离差分数(deviation score)	四分位数(quartile)
离差(dispersion)	全距(range)
四分位距(interquartile range)	抽样误差(samping error)
平均数(mean)	标准差(standard deviation)
平均差(mean deviation)	统计量(statistic)
中数(median)	方差(variance)
众数(mode)	

8.5 分组数据——分布

我们在一项研究中所收集的数据被称为数据组。如果你对一个十二人组成的样本进行一个简单的实验，你将获得一组包括 12 个数据的数据组，并且对于数据，除了用平均数和标准差来总结它之外，你真的没有其他更多可以做的。然而想象一下，你对整个学院或学校的学生进行一个调查，让他们评价饮食设施的质量。当拥有像这样的大量数据时我们就能够看到整个数据的分布。

表 8.1 表明一种分类或称名变量结果——"个体类型"及记录的该类型人数。这个表示人数的数字被称为发生事件的"频数"。这些频率显示于表中的第二列，整个表就被称为**频数分布**(frequency distribution)。假定我们记录了父母首次注意到他们的孩子说出电报式语言的月份(如"妈妈袜子")，如果记录了许多父母的数据，将不能用表呈现所有这些数据，但我们能像在表 8.8 记录的那样，表示出所有数据的分布。

表 8.8 父母首次注意的电报式言语年龄

年龄(月份)	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	总计
报告的孩子数量(频数分布)	1	0	5	12	37	64	59	83	17	41	12	0	4	5	0	340

带有分组区间的频数表

如果我们测量某个变量，如时间或重量，我们必须处理好测量区间。我们在前面的遇到过这个问题。当我们测量人们猜出字谜所花费的时间不能测出精确的秒数时，如果用秒数来计量，停表可以精确测到 0.1 秒，那么我们可以测量到最接近的秒数。在 17.5 到 18.4 之间的任何时间都被计作 18 秒。在表格 8.9 中可以注意到：**组距**(class intervals)是"10～15"，"15～20"，等等，这种表示通常出现在拙劣的表格中。因为这样表示我们将会遇到一个问题，即 15 应归入哪一分类中？15～19 的区间不包括 19.5 及其以上的任何数据。由于我们的计时器至少能测量到 0.1 秒，因此能有效地测量 19.4 或以下的任何数据。19.5 或 19.6 秒的时间将归入 20～24 的区间内。这是因为我们不能精确测量到 20 秒，我们能测量到的是 19.5～20.4 秒这个时间段。

表 8.9　猜谜时间(秒)——分类区间内的频数

秒　数	参与人数	累积频数	参与人数少于
5~9	2	2	9.5
10~14	0	2	14.5
15~19	8	10	19.5
20~24	18	28	24.5
25~29	13	41	29.5
30~34	5	46	34.5
34~39	4	50	39.5

带组距的频数表格。

表 8.9 表明属于每个组距的参与者的频数(f)。因此,我们能看出有 8 个参与者花了 15~19 秒的时间来猜出谜语。表格中的下一列是**累积频数**(cumulative frequency),也就是,得分少于这个区间上限的人数。就像上面解释的那样,这个区间的上限实际上为 19.5。对于 15~19 的区间,累积频数是在该区间的频数 8 基础上加上低于它的 2,因此为 10。

8.6　图形表述

对于读者来说,用图表的方式来考察数据的分布情况是非常有用的。在你写一份报告时,注意到这些图表被称为图示,正如其他任何图画条目。表格是不一样的:它有列和行,被称为列表。这里有一些相关提示,在呈现任何图表时都需要铭记在心。

- **不要在优美和多样上浪费时间**。绘制图表不是一个如何使图表美观和给人以深刻印象的竞赛。你只需要呈现符合要求的、清晰的信息。像 Excel 这样的程序将会产生非常复杂的各色三维图表,但如果你使用不小心,它将会分散阅读者的注意力。坚持一个简单的、常规的原则,在任何情况下都不要仅仅为了花哨而用大量不同的图表来呈现相同的数据。
- **不要描绘原始数据**。图表只是数据的一个概括。读者希望看到的是反映数据整体情况的简洁图像。常见的错误是产生一个像图 8.7 一样的图表,用条形图或列表来代表每个参与者的得分。这里我们看到,图表没有给出数据的整体概况。它不是对数据的概括,并且非常不规范。条形图仅仅是按着参与者被测试的任意顺序安排的。在图示中,我们不期望任何模式并且我们也没有任何模式。显然这个图表是没有任何意义的。

条形图

条形图通常表示分组或分类,但是在横轴上的分类也可以是报纸、年份、商店。实际上,对于有统计价值的每一种类都可以加以系列分类。横轴或 X 轴通常描述称名或分类变量,即使它是像图 8.8 中的年份。这里选出的年份是为了表明作者认为重要的对比。如果所有有价值的年份都被表示出来,那么图表将会变成一个特殊的条形图——直方图。因为 X 轴代表谨慎的分类,条形图的条块应该总是分开的。

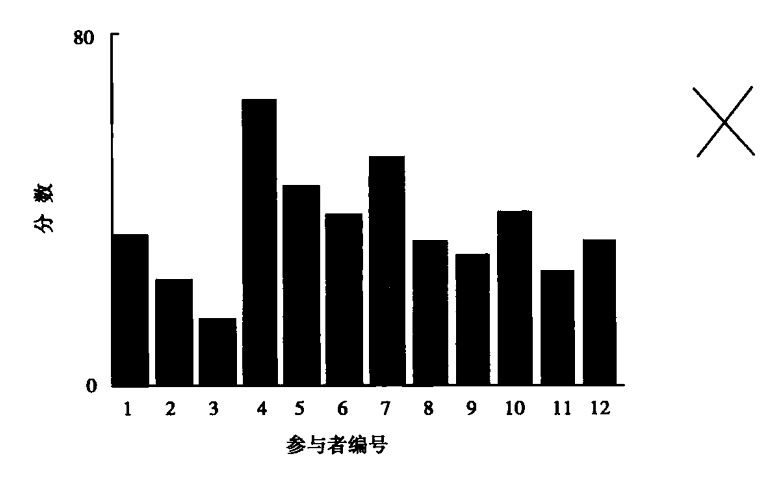


图 8.7 关于个体参与者分数的不合理图表

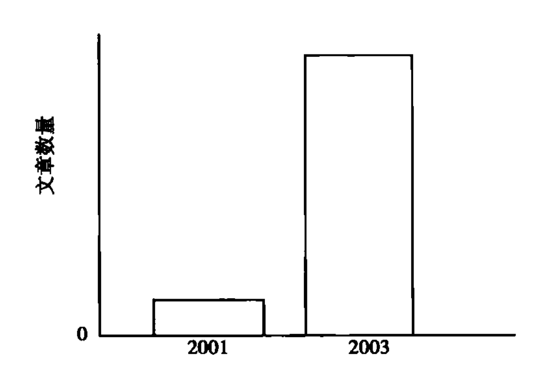


图 8.8 已出版的有关恐怖主义的心理学文章的数量

正确的条形图的例子如图 8.9(a)所示。我们能看到2004 年女生在数学成绩 A 水平上得分为 A 到 C 的数量稍稍多于男生。实际的百分比是女生 79.1% 和男生 73.6%。然而,一个不择手段的新闻记者为了一篇题目为“女孩在数学上领先”或“对男生在 A 水平数学方面的能力下降的高度关注”的文章,像图 8.9(b)那样夸大这种差异。这个图表是一个误导,因为它没有澄清一个事实:即 Y 轴(左边纵轴)没有以零为起点。因此,读者的注意力都吸引到男生和女生结果间的不真实的“巨大”差异上。当然,两者间有不同,但差距并不像图 8.9(b)表示的那样大。

如果 Y 轴从零点开始,纵列的高度将会变得不雅观,图表中 Y 轴明显的中断是为了告诉读者不是所有的 Y 轴都像图 8.10 表示的那样。这个条形图也是联合图表的一个例子。在联合图表中,两种结果一起被呈现,这些用不同的彼此接近的带阴影的条块来表示。

直方图

表格 8.9 中数据的直方图将像图 8.11 显示的图表那样,每一个纵列的宽度必须相等并且代表组距的大小。区间利用它的中值来表示。第一个区间(如左边表示)是 5 ~9。正如前面解释的那样,这个区间的上下限是从 4.5 ~9.5,因此这个区间精确的中值为 7。每一列的高度表示(并且仅能表示)落入该区间的事件的数量

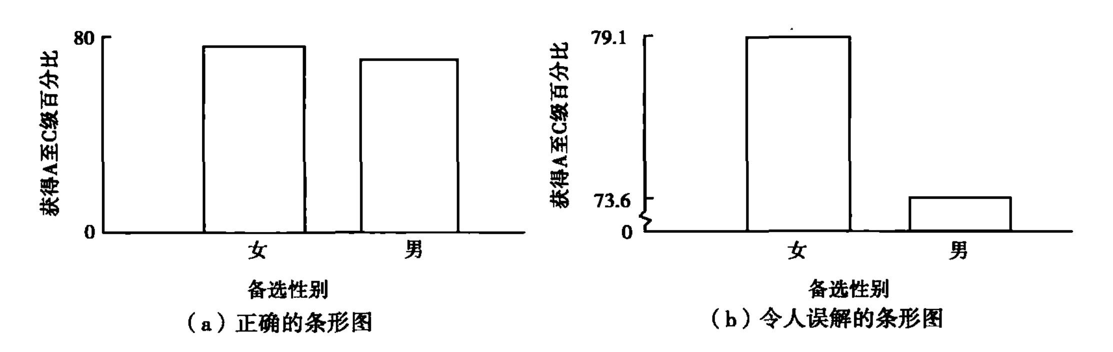


图8.9 2004年夏在数学A水平方面,女生和男生得分为A到C等级的百分比(所有的版面)
来源:《卫报》-http//education. guardian. co. uk/alevels2004/story/0,14505,1285751,00. html

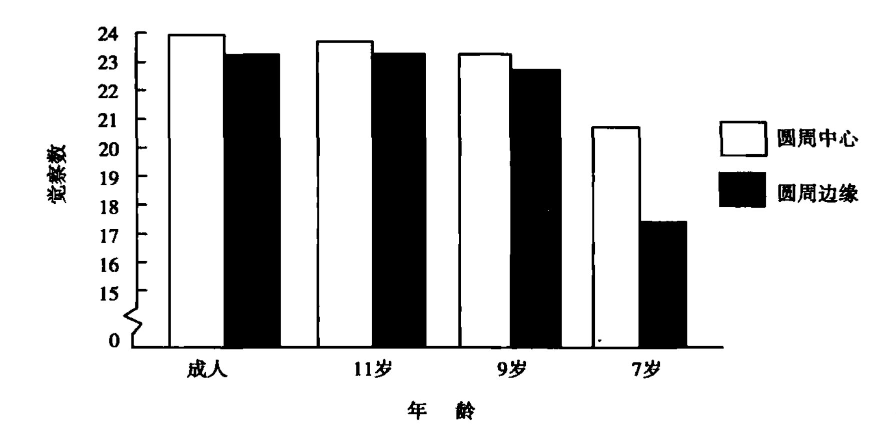


图8.10 四个年龄组在对圆周中心和圆周边缘明显运动觉察的平均数目
来源:David,Chapman,Foot and Sheehy,1986;英国心理学杂志(英国心理学学会授权下再版)

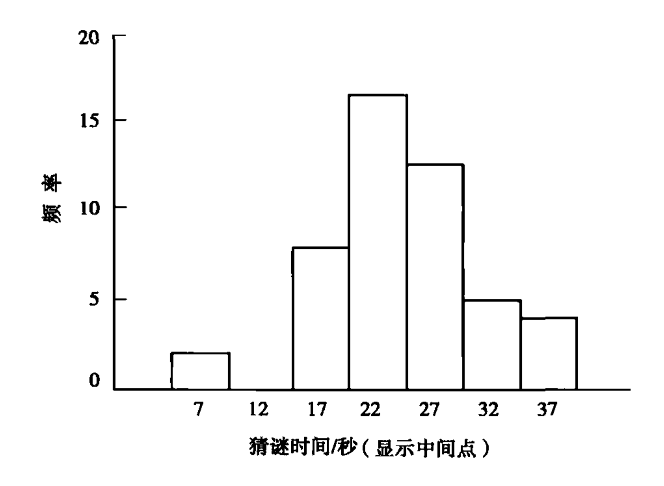


图8.11 猜谜用时(用区间表示)频数

(频数)。因此15~19的区间(中值为17)的纵列有8个单位的高度,因为表格8.9告诉我们在这个区间有8个人。注意,所有的条块都组合在一起,并且如果在一个区间内没有任何事件,我们需要留一个齿形的间隙,并假设两边都有事件。这样做的原因是,一旦我们在短时间内看一个正态分布时,这个直方图将会变得更清楚。

从根本上说，任何条块的面积或一组条块都是与表示的事件的数量成比例。因此，由于条块有 1 个单位的宽度，所以区间 5～9 有两个事件的高度的条块，面积为 2。由于总共有 50 个事件，直方图的总面积为 50，区间 5～9 占据了整个图表面积的 2/50。由于从 10～14 没有任何事件，我们必须表明它是没有面积的(即没有事件)的。你能说 5～9 区间占了整个面积的 4%，因为 2 是整个面积 50 的 4%。然而，习惯上通常把图表上的整个面积当做一个单位，即单位 1。然后用小数表示其中的部分。这就意味着 5～9 区间占了 0.04 个面积，和 4% 一样，如果你对小数有点迷糊，表格 8.10 中的分数和百分数可能会有帮助。

我们通常把区间或分数放在 X 轴(水平的一个)，把频率放在 Y 轴(垂直的一个)。一些程序(如 Excel)可以表示一个直方图，在直方图左边表示频数，条块在水平方向按从小到大的顺序由左向右排列。

表 8.10 承认它吧！你已经忘记了如何把分数转变为小数，或者诸如此类的问题

百分数→分数→小数

从 5% 开始，移走百分号%，把数字放在 100 上面 $\left(=\frac{5}{100}\right)$，用 5 除以 100，在 5 的右边加上一个小数点，并且把小数点左移两位：

$\frac{5}{100}\rightarrow 5.0\rightarrow 0.5\rightarrow 0.05$

如果已经有小数点，只需要正确的移动：

$2.5\%\rightarrow\frac{2.5}{100}\rightarrow 0.25\rightarrow 0.025$

小数→分数→百分数

从 0.01 开始

- 小数→分数

任何小数都可以按如下步骤变成分数：

- ❑ 在小数部分有几个数，就在 1 后相应地加几个 0。
- ❑ 去掉第一个数字左边的所有 0。

  0.01→小数点后有两位数字，所以→$\frac{1}{100}$

  0.375→小数点后有三位数字，所以→$\frac{375}{1\,000}$

- **分数→百分数**

百分数只是以 100 为分母的分数，所以 $\frac{1}{100}=1\%$，$\frac{43}{100}=43\%$

所以 $0.01\rightarrow\frac{1}{100}\rightarrow 1\%$

对于分数 $\frac{375}{1\,000}$，需要把分母变为 100，通过分子、分母同时除以 10 来完成这种转换

$\frac{375}{1\,000}\rightarrow\frac{37.5}{100}$(除以 10 只是把小数点向左移一位，所以 375 变成 37.5)

因此，$0.375\rightarrow\frac{375}{1\,000}\rightarrow 37.5\%$

正态分布

在这一章的前面我们了解到对于一个区间范围的测量应该真正被看做区间而不是一个确切的值。想想测量某人的身高，在挤压头发和为短袜的厚度以及踮脚尖争论后，最后你能做的是看你卷尺尺寸上的最接近的一条线。某人的高度看上去刚好在线上是非常少见的。实际上，你能做的是把他们放在一个区间内——比起 164 cm 更接近 163 cm，因此高度为 162.5 ~ 163.5 cm。如果我们以大量人群作为样本，用这种方式测量他们的高度，然后把它们的高度绘制在一个直方图上，我们将得到一个看上去很美观的图表，如图 8.12 所示。因为统计学家要对由条状物形成的曲线下的面积做大量的工作，我们必须把纵列绘制为一个紧挨着一个，它不像条形图，是没有间隙的。

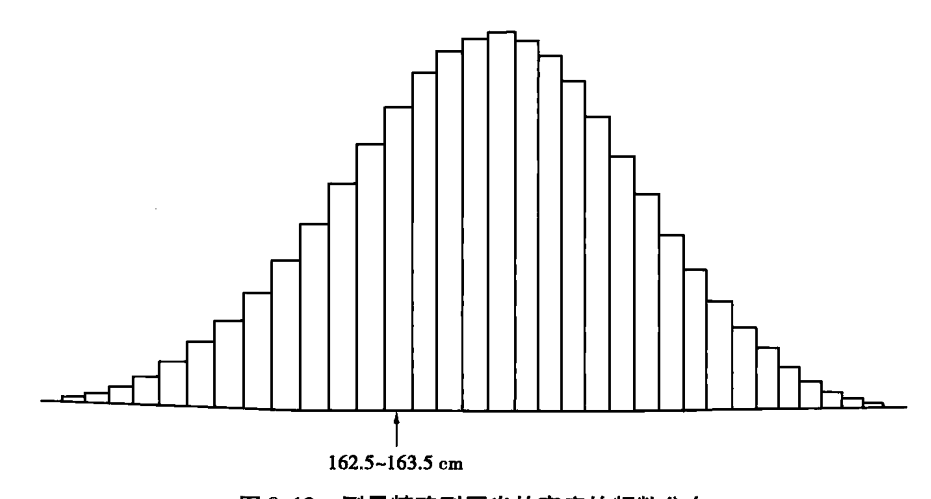


图 8.12 测量精确到厘米的高度的频数分布

这个直方图的形状近似于一个非常重要的数学曲线——钟形曲线，参见图 8.13，这是我希望的相当明显的理由。它同样更正式地被称为**正态分布**（normal diatribution）。记住，没有一列数据完全符合一个正态分布——那是随机选取事件的一种方式，但是每一个随机大样本的分布都将近似于这个分布——你也能期望它是这样。

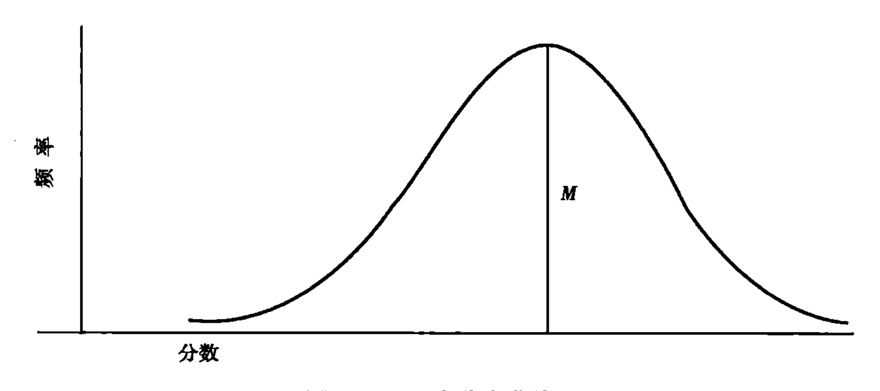


图 8.13 正态分布曲线

近似正态曲线和常态人群

记住下面的内容非常重要，当心理变量被说成“正态分布”或“符合正态分布标准”时，我们所讨论的一直都是一个近似的完全正态曲线。选取的被试群体与理想群体总会有所差异。之所以出现这一问题，是因为当我们进行差异检验时，统计理论往往假设所选取的样本群体符合正态分布。如果总体变量值不符合正态分布，差异检验得出的结论可能出现严重的错误。“正态”这一术语被用来描述人群的数量分布。曲线之所以被称为“正态”完全是由于数学的原因（你可能记得“正态”一词就如同在几何中使用“垂直”一词一样）。

正态分布曲线[①]下的面积

假设我们设计一个针对8岁儿童的阅读测验，用它来测试一个大样本的儿童，并找到平均数和标准差。无论原始数据怎样，都能对其加以转换，使平均数和标准差转变成图形以便于处理。这一程序是前面第6章提到的标准化过程的一部分。其余过程涉及对测验加以处理直到使它确实产生一个正态分布为止。假设一个代表性的8岁儿童样本的平均数是40，标准差是10，显而易见，50%的8岁儿童得分在40以上，另外50%则在40以下。顶端50%的面积是图8.14中所有的阴影部分。

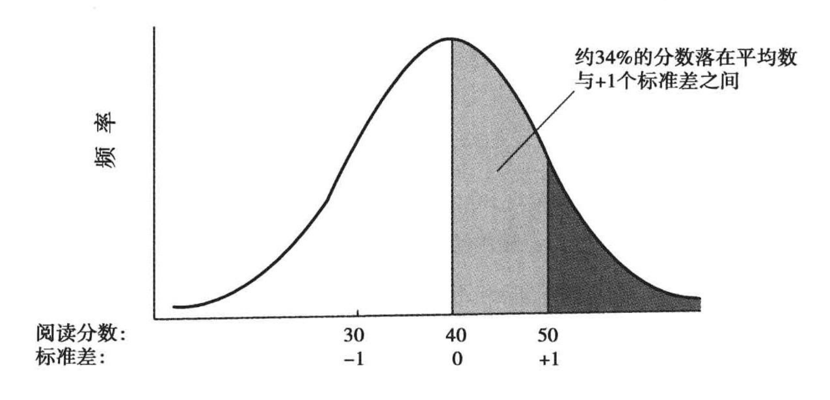


图8.14 8岁儿童的阅读分数分布

记住前面我们在统计运算时使用的标准差。它是所有数据对平均数的一种平均离散程度。统计学中，如果一种分布是正态的，那么我们就知道标准差处于分布曲线的什么位置；也就是说，我们知道标准差以上和以下的百分数。正因为如此，研究人员才花费大量的努力，试图使心理测量符合正态分布。

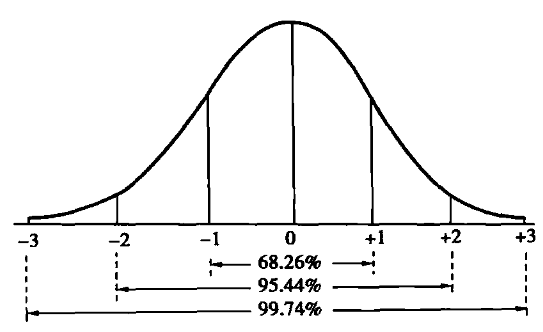
从图8.14中可以看到，约34%的分数介于平均数和平均数以上1个标准差之间。这并不奇怪！因为它是正态分布的。因此，大约34%的8岁儿童的分数为40～50，这是因为平均数是40并且标准差是10。同样的数字（34%）也适用于那些在平均数以下1个标准差的儿童数量。

更详细的内容如下：

① 本文中的正态分布曲线是指标准正态分布曲线。——译者注

百分数在平均数和 ±1 个标准差之间	34.13
百分数在平均数和 ±2 个标准差之间	47.72
百分数在平均数和 ±3 个标准差之间	49.87

以上情形如图 8.15 所示，但在这里我们把百分率增加了一倍来显示曲线的面积，因此，百分数在平均数以上和以下 1、2 或 3 个标准差之间。例如，95.44% 的分数应介于 -2 个标准差和 +2 个标准差之间。



正态曲线中 $-n$ 与 $+n$ 个标准差之间的面积

图 8.15　标准差位置与正态分布下相应的面积

标准分数（Z 分数）和正态分布

如果一个人的得分恰好是平均数以上 1 个标准差，他的 Z 分数就是 +1（也称为标准分）。假设全国的平均智商为 100，标准差为 15，如果你在 IQ 上的得分是 115，你的 Z 分数就是 +1。我们可以根据平均数和标准差计算 Z 分数，比如一个智商为 130 的人，其 Z 分数就是 +2。

一般性定义：Z 分数就是一个数高于或低于平均数有几个标准差

让我们用一个简单的例子来介绍 Z 分数的计算公式。假设你们班级里鞋子的平均尺寸是 7，标准差是 2，但大脚哈瑞的鞋子尺寸是 11。那么他高于平均数几个标准差？也许你很容易就能够计算出，但让我们看看你是如何计算的。你需要以下数据：

$$x = 11 \qquad \bar{x} = 7 \qquad s = 2$$

你肯定会这么想："哈瑞比平均数高 4（11 - 7），因此离差是 4 个尺寸；每个标准差是 2，所以他高于平均数 2 个标准差，即他的 Z 分数是 2"。基本上，你是用哈瑞的离差除以他的标准差，下面的公式说明了你的计算过程：

$$z = \frac{x - \bar{x}}{s}$$

这里仅仅是用离差除以标准差。Z 分数的计算通常并不像我所计算的那样友好，平均数和标准差很少是这些简单的数字。但是你必须记住的是，Z 分数的计算需要用离差除以标准差。你可以根据原始分数计算出的平均数求得离差。

如果回头看看图 8.15,你可以看到正态分布上高于哈瑞的人并不多。假设鞋子的尺寸就像我们期望的那样符合正态分布,大概有 2.5% 的人高于哈瑞。为什么? 由于图 8.15 中有大约 95% 的人(具体是 95.44%)介于 -2 和 +2 个标准差之间。有 5% 的人在此范围之外,其中一半处于非常低的位置,另一半则处于最上方。因此有 2.5% 的人高于哈瑞。

我希望上述计算不是太难的工作。但是,我们如何计算 Z 值为 1.5 的百分率呢? 这在我们的图中没有体现。幸运的是,已经有人为我们提供了答案。翻到附录 2 中的表 2,你可以看到 1.5 个 Z 分数出现在靠近偏左一列的底部。主要有三列,第一列是 Z 值,第二列是切除高于平均数的曲线的面积数;第三列是左侧高于 Z 值的数量(即右边的曲线)。这里的面积就是图解中第二和第三列顶部的阴影部分。在这个例子中我们发现,曲线的 0.433 2(即 43.32%)介于平均数和 1.5 个 Z 分数之间。高于 1.5 个 Z 分数的有 6.68%,可以从第三个柱子中看到这一结果。

Z 分数为负值的计算方法如同使用一面镜子。-2.2 个 Z 分数对应的面积为 0.486 1,其左侧的平均数是 0.486 ,即 48.6%。平均数的整个左侧曲线下的面积是 50%,然后再从 50 中减去 48.6 ,我们发现低于左侧 -2.2 个标准差的只有 1.4%。

心理量表、标准化与正态分布

Z 分数和正态曲线下面积之间的关系在心理测验领域是至关重要的。如果(并且是一个很大的"如果")一个变量被假定为在人群中是正态分布的,并且有一个进行标准化测验的大样本,那么,我们就可以通过把原始分数转化为 Z 分数,来快速评估某一群体的相对位置。这对于评估来说是宝贵的,例如,评估儿童的阅读能力、一般智力或语言发展、成人压力、焦虑、职业倾向(采访中)等。教育心理学家可以告诉你,相对于其他年龄阶段的儿童,这些儿童有哪些不同。这里你也应当指出,IQ 测验并不能测出自然的智力正态分布。而是有意识地将 IQ 测验符合正态分布,这基本上是服务于一定的研究目的以及便于比较测试结果的方便。通常,一个 IQ 测验是标准化测验(原始分数被调整过),其平均数是 100,标准差是 15。

偏态分布

如果一个分布在一个方向比另一个方向有更多的极端分数,就称之为是该方向上的偏态,见图 8.16。如果偏态过大则导致平均数的严重失真,而其他测量则相对地代表着集中趋势。反应时间往往是正偏态,因为它很容易比平均时间慢却很难快于平均时间。

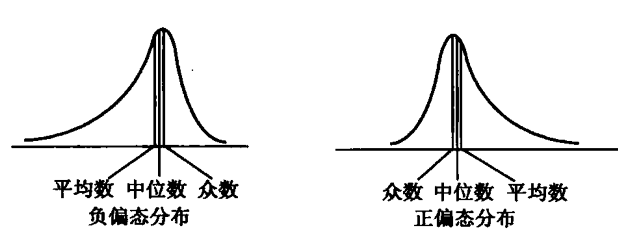


图 8.16 偏态分布

练 习

1. 为下列数据找出合适的组距并画出直方图：

 62 65 71 72 73 75 76 77 79 80 82 83 92 100 106 117 127

 65 70 72 72 74 75 76 77 79 80 82 88 93 102 110 121 128

 65 70 72 73 74 76 76 78 80 81 83 90 95 103 112 122 135

 (a)检查直方图，并为测量选一个恰当的集中趋势量。

 (b)计算该集中趋势量。

2. (a)粗略描绘两个平均数相同但标准差不同的正态分布图。

 (b)描绘两个标准差相同但平均数不同的正态分布图。

3. 智商分数分布的平均数是100，标准差是15：

 (a)16%以上的人的智商值是多少？

 (b)分数低于70的百分率是多少？

 (c)分数低于-1.3个Z分数的百分率是多少？

 (d)Z值为2的人的智商是多少？

答 案

1. (a)数据是偏态分布，因此用中位数

 (b)79。

2. 见图8.17

3. (a)115；(b)约2.3%；(c)9.68%；(d)130

关键术语

条形图(bar chart)

组距(class interval)

累积频率(cumulative frequency)

数据集(dataset)

频数分布(frequency distribution)

标准分数(Z分数)(standard score(z score))

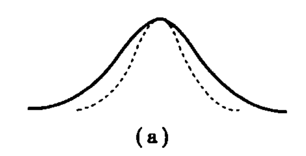
直方图(histogram)

负偏态(negative skew)

正态分布(normal distribution)

正偏态(positive skew)

偏态分布(skewed distribution)


(a)

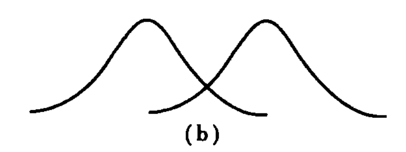

(b)

图8.17 上述《练习》中问题2的可能答案

8.7 统计注释与符号

N——样本量

N_a——样本A的数量

x——样本中的数值，如某人的分数

y——另一个测量变量的数值

$\sum$ ——希腊字母 S（“sigma”）——意思是“把每一个数值相加”。例如：

$\sum x$ 是指“把样本中所有的 x 加在一起”。（进一步阐释如下）

统计符号

样　本			总　体		
平均数	标准差	方差	平均数	标准差	方差
$\bar{x}$	s	s^2	μ	σ	σ^2

注：“M”和“SD”用于结果报告。见专栏 11.1。

一些规则

1. 数学公式容易令人混淆，尤其是在统计中，使用乘号（×）是因为运算过程中有太多的“x”。“x”是指一个特定分数或值。公式中的乘号通常省略，因此当一个值紧接着另一个值时，我们需要把它们乘起来。例如，rN 意思是“用 N 乘以 r”。
2. 先进行括号内运算，或进行过 $\sum$ 或 $\sqrt{\ }$ 运算后，再进行其他运算。

以下是关于运算规则的一些例子：

$\sum xy$—— 用所有的 x 乘以所有相应的 y 然后把所有的结果相加。注意 xy 是指“用 y 乘以 x”并且只有在完成所有的 x 和 y 乘法运算后才能把结果相加。

$\sum x^2$—— 先计算所有 x 的平方，再把它们相加。注意与下面的符号加以区别。

$(\sum x)^2$—— 把所有的 x 相加，然后再加以平方。

$\sum x \sum y$—— 用所有 y 的和乘以所有 x 的和。

$\sum (x-\bar{x})^2$——用每一个分数 x 减去平均数，求出它们的平方，然后再把所有的结果相加。

$(N-1)(N-2)$——求出 $N-1$ 和 $N-2$，再把二者的结果相乘。

$r\sqrt{\left(\frac{(N-2)}{(1-r^2)}\right)}$的运算步骤是：

(1)求出 r^2

(2)求出 $1-r^2$

(3)求出 $N-2$

(4)用第三步除以第二步

(5)求出第四步的平方根

(6)用 r 乘以第五步的结果

9

显著性检验和概率

本章内容

- 本章内容涉及如何对以下问题作出判断，即在一项研究中显示的效应（差异或相关）可能是一个真正的效应，而不仅仅是在数据上的随机波动。
- 我们将考察显著性检验的基本原理。
- 我们将介绍虚无假设和备择或实验假设的概念。
- 给出了概率的简介。我们需要理解拒绝虚无假设的依据（当在虚无假设情况下结果发生的概率太低时）。
- 对显著性水平的描述，主要用0.05，但也用0.01 和0.1。
- 介绍更复杂的概念，即Ⅰ型错误和Ⅱ型错误，以及定向假设与单侧检验和非定向假设与双侧检验之间的关系。

我们回头看一看第8章的专栏8.1到专栏8.4。我们介绍了一个在日常生活中极其普遍的情景，即你不得不作出判断：差异是由一些真正的效应引起的，还是因“纯粹的巧合”而可以被忽略的。本章我们主要讨论这些问题，但在这里我们以正式的和精确的方式来做。在以下章节的每次推断统计检验(inferential statistical test)中，通常需要对以下两个问题作出判断：

1. 这些数据能够提供令人信服的有效应的证据吗(正在发生的)？
2. 这些数据能看作仅仅是一些随机因素的作用吗？

让我们看一个心理学中的实际例子。最好记住，在进行显著性检验(significance test)时我们正在进行科学研究，需要检验我们的证据是否支持来自假设的预言。一项研究报告的论点是这样形成的：

观察和论点　当人们被观察着去完成一个简单任务时，他们常常会尽力做好，其完成任务的行为表现趋于改善。特里普利特(Triplett,1898)第一个进行该项研究。他观察到自行车手与他人一起练习时比单独练习时更努力。然而在更为复杂的任务中，人们常常抱怨他们被观察，并说这使他们分心。他们看来也似乎完成得比较糟糕。

待检验的假说　在复杂任务中，人们在被观察时的表现要比其单独时的表现更糟。

研究和预测　被试将完成“环绕线”的任务，就像在乡村节日里看见的那样。他们必须使捆在一根棍子的末端的一圈电线通过一个通电的环线，而不接触它。接触导线的次数将被记录。他们先由一个人单独完成，然后在有六名观众观察的情况下来做。各种条件将被相互平衡抵消。我们预测在观察条件下被试的接触导线平均次数将比单独条件下多。

这里我们有一个假说和一个以此假设拟定的实证性检验(empirical test)，伴随这个的是一个具有操作性的研究预测(research prediction)。注意：如果发现研究预测是有效的，它将支持这个研究假设。让我们假设有两组学生来进行实验，一组学生对于完成该项实验的任何事情都非常谨慎仔细，而另一组学生却是马虎、粗心的。让我们来看表9.1中显示的结果。

马虎学生组的数据表明实验起作用了，因为单独条件下的平均数比观察条件下的平均数小。这个表明是观众引起人们糟糕表现的证据能令人感到信服吗？当然，尤其是当你看到在每种条件下，其内部所用时间的变化时(在观察条件下，从1.3秒到16.2秒，相差15秒)差异也相当的小。在两个平均数之间仅仅存在0.74秒的差异看起来并不是那么令人信服，不是吗？

然而，当看到态度认真的学生组得出的结果时，我们得到了平均数之间大约是8秒的差异，即使这种离差(dispersion)与态度马虎学生的结果相似。在心理科学研究中，有些结果可能看似极其相似，但我们不能依靠直觉来说这个结果“看上去是对的”。我们需要一个协定来告诉我们，什么时候结果推断出了一个真正的潜在差异(即“效应”)，什么时候推断结果并不显著地高于随机水平。而这样的协议应让

每个涉及研究的人都能同意并理解。接下来让我们一起来看这个检测显著性的系统。

表9.1 两组学生在"环绕线"任务里,被试"单独"和"在观察者面前"接触环绕线的时间(秒)

(a)态度马虎学生组的结果		(b)态度认真学生组的结果	
有观察者	**单独**	**有观察者**	**单独**
9.9	5.2	14.8	4.1
15.1	3.3	13.5	3.2
16.2	9.1	9.6	5.7
11.7	6.1	8.7	9.0
11.1	16.2	12.2	11.3
9.1	2.7	15.3	2.8
10.0	7.8	18.6	5.1
8.5	18.2	12.1	6.1
1.3	10.3	15.8	3.3
2.4	9.0	17.1	7.2
平均数=9.53	平均数=8.79	平均数=13.77	平均数=5.78

9.1 显著性检验中的虚无假设和备择假设

人们发现得出显著性概念很难,但是显著性检验所需的一切逻辑都包含在以下阿尔菲的思维中:

> 杰克:"我学会怎么掷硬币了——你看,一个正面向上,两个正面向上……连续八个正面向上。"
>
> 阿尔菲:"让我看看那个硬币。你用普通的硬币不可能常常有八个正面向上,我敢打赌这个硬币是固定了的。"

在日常生活中我们常常使用到的显著性检验的逻辑就同此一样简单。在我们推断送奶工是否对我们要价过高的时候,在我们推断右车道是否行驶速度更快的时候,都会使用到显著性检验的逻辑。

阿尔菲思考的实质如下:

- 假定硬币是公正的或"公平的"。
- 如果硬币是公正的,估计连续出现八个正面向上的概率(probability)。
- 如果概率非常低,然后推断硬币是不公正的。

这种最初的假设被称为做**虚无假设**(null hypothesis)。这里硬币是公正的,或者从形式上来说,在掷硬币的次数非常多的情况下,硬币落地是正面和反面的频率是相等的。这种假设的形式使阿尔菲估计得到连续八个正面向上的概率非常小——小到只能当做是巧合。因此他拒绝了用这种巧合来解释投掷硬币的结果,并假设硬币是不公正的。这种假设被称为**备择假设**(alternative hypothesis)。备择假设是虚无假设的反面,也叫**实验假设**(experimental hypothesis),通常是我们的研究一开始就支持的。我们把虚无假设和备择假设标记为:

虚无假设　　H_0：硬币出现正面向上和反面向上的次数相同。

备择假设　　H_1：硬币出现正面向上和反面向上的次数不相同。

为什么我们需要虚无假设

我们需要虚无假设是由于我们不能简单地反复说："'偶然'发生的概率有多大？我们用'偶然'来表示什么？我们用'偶然'来表示'如果什么都不会发生'或'如果没有受到其他的影响'。当学习显著性时，你应当经常尝试用"如果虚无假设是正确的"来替换"偶然"。所以，你应明确知道"若虚无假设是正确的，其发生的概率"，而不是模糊地知道"其偶然发生的概率"。

以上我们介绍的形式在日常交谈中得到反映。我们说的事情，类似"如果他们没有相互见面，他怎么会知道她的号码"或"如果送奶工没有利欲熏心，怎么会出现他常常少给我们一品脱[①]的牛奶"。

这里列出了一些关于虚无假设的重要说明：

1. 虚无假设往往是对世界上事情状态的描述。
2. 它是统计学上对于总体的描述。
3. 它不是预言将要发生什么事情，而是假设事情是什么。
4. 虚无假设是对于在通常情况下没有效应的一种假设。

让我们将"日常生活"中的一个例子展开，更加清楚地阐述虚无假设的作用。

在手套销售部的工作——学习虚无假设的理想一课

想象你正在百货商店内的一个手套销售部工作。你下仓库去找一双特殊的手套。你知道，至少你认为自己知道，抽屉里有许多左手手套和右手手套并且数量上是相同的。突然停电了，你处于黑暗中。你想："没问题的，如果我取出几个手套就一定能匹配成一副的。"假定在你的取样里，你可以拿到足够多的左手手套和右手手套，你取出五只。然而当供电重新恢复时，你发现取出的都是右手手套。你可能会把这当成是明显的巧合而毫不在意，如果没有人动过抽屉的话这肯定就是巧合了。但从另一方面来讲，如果是你的某位同事想开个玩笑或非常健忘，你可能会怀疑他们摆弄过放手套的抽屉。你为什么会有这样的怀疑呢？



图 9.1　随机抽取到五只右手手套的可能性是多少

答案是：因为如果仅仅一半的手套的确是右手的，那么抽取五只相同的手套是几乎不可能的结果。让我们用阿尔菲的逻辑分析以上现象：

1. 设定虚无假设——H_0：左手手套和右手手套数量相同（注意这是对在抽屉里手套

① 液体容量单位，等于八分之一加仑。——译者注

"总数"的设定)。

2. 计算抽取五只均为右手手套的概率。
3. 如果概率非常低,则明显是巧合。假设抽屉里左手手套和右手手套的数量并不相同——这就是备择假设(H_1)。

更正式地说,这些是显著性检验必须考虑的三个步骤。

表达显著性的语言

如果结果对巧合来说概率太低,我们会说有一个显著性的结果,我们选取了数量显著的右手手套。我们说在"单独"和"观察"条件下的平均误差的差异是显著的。

9.2 概 率

上面的第二步需要一点扩展。如果虚无假设是正确的,我们如何计算连续选择五只手套的概率?虚无假设假定:左手手套和右手手套的数量相同。选取右手手套的概率应该是 $\frac{1}{2}$ 。为什么呢?

当我们计算可能性相等的事件的概率时,如计算掷一枚硬币出现正面向上或投一个骰子出现"6"的概率,我们用以下公式:

$$p = \frac{\text{某个特定结果出现的次数}}{\text{可能出现结果的总次数}}$$

用硬币时有两个可能的结果,正面向上或反面向上,因此 2 用在公式下面。我们关心的是正面向上,这个仅在两次结果中出现一次。因此正面向上的概率是$\frac{1}{2}$。你可能会想为什么连续掷正面向上的概率是$\frac{1}{4}$?请看专栏 9.1,你在专栏 9.1 里的

专栏9.1 对概率赋值

> 请看以下叙述,你会发现你对这些事件中的大部分发生的可能性都会进行设想。依据你认为发生的可能性,对每个事件的叙述在0(一点也不可能)和100(完全可能)之间赋值。
>
> 1. 下周三会下雨。
> 2. 下月的第一天你会吃早饭。
> 3. 你的心理学老师在下节课会打喷嚏。
> 4. 明天一早太阳会升起。
> 5. 今天晚些时候你会想到大象。
> 6. 公正地掷骰子会出现一个 5 或 6。
> 7. 公正地掷两枚硬币会出现两个反面向上。

第7项将看到连续掷正面向上的概率是$\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$。如果虚无假设是成立的，这个概率同从抽屉中取出两只右手手套的概率是一样的。[①]

概率通常标记为p，在这个范围上以小数赋值：

0 ←——————to——————→ 1

不可能 必然发生

如果你对每个事件在1～100之间进行赋值，把他们平均分成100份，他们将在以上的范围内上出现小数值。对于第一条，如果你住在英国，无论何时你回答了50，在这个范围上就变成了0.5。然而，如果你住在孟买，在10月你可能会回答5，在范围上就变成了0.05。在显著性检验中你需要更加小心，因为某些人忘记了这里是小数，就相差大了，注意0.05比0.5小十倍。对于第四条，如果明天太阳不会升起是极小可能的，不能赋值1（这样的举例在心理学研究方法中是无用的!）。对于第五条在你读了之后可能已经赋值了。对于第二条有赖于你的早餐习惯。

对第六和第七条我们可以应用以上的公式。第六条有6种可能出现的结果，而我们仅仅关注这其中的两个结果，因此可能性是$\frac{2}{6}$，简化为$\frac{1}{3}$，或是0.33。第七条有四个可能出现的结果（正面和正面，正面和反面，反面和正面，反面和反面），我们关注的是其中一个，因此p值是$\frac{1}{4}$。

专栏9.2 注意主观性概率——大雨并不是你能左右的

> 我们必须把日常的主观评估放到我们的考虑中，并依靠精确的数据和计算来研究概率。在暑假，我们经常诅咒这样的情形：当我割完草，刚搬走孩子的自行车、足球、弹簧棒、日光浴设备时，天就开始下雨了。同样的事情也发生在超市里，当人们排队付款——一些人经常有5个未标价的物品，或者有许多购物优惠券要处理，或16岁的店员把茶叶弄撒了，或微笑的监管者想换一下松了的袋子。由于一些人想右拐弯，使我在行车道上不得不放慢行驶速度——不是这样吗？我不得不承认这就好像某天我刚割完草坪就开始下雨一样。也许是我只记住了那些糟糕的日子了吧？

我们可以由此得出结果，第一个正面的概率是$\frac{1}{2}$，第二个正面的概率也是$\frac{1}{2}$。把事件1出现的概率和事件2出现的概率相乘，$\frac{1}{2}\times\frac{1}{2}$，等于$\frac{1}{4}$。

回到手套问题

我们现在回到关于手套的问题中，若虚无假设是正确的，对随机抽取五只手套的概率进行赋值。注意我们很容易说：我们计算"在H_0成立条件下结果出现的概

① 数学家们说，既然我们是拿走一个手套而不是替换它，那么，这就不完全正确了。因此，从我们取出第二只手套开始，抽屉里的左手和右手的手套的比例不再是50∶50了。事实的确是这样，但我假设有足够多的手套，以致那些未归还的手套仅仅对概率的结果产生细微的影响——就让我们使事情变得更简单些吧。

率”,是$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{32}=0.03$(忘记了可替代的问题)。现在,我们问,0.03是否足够低到以至于可以假定虚无假设是不正确的呢?换一种说法,0.03是否足够低到以至于可以假定有一些人动过抽屉了呢?那样的推断有赖于其他研究者认为什么水平是拒绝偶然性的合适水平,我们应该在那样的水平上作出判断。要知道,0.03不是结果“随机出现”的概率,它不是“虚无假设成立的概率”①。

如果虚无假设成立,这是结果出现的概率。

这就是我们对p值的定义。

我们将这点概括为:

- 我们已经从假设有左手手套和右手手套的抽屉中选取五只手套样品。
- 我们觉得选五只右手手套的情况不经常发生,可能事实上有人动过了装有手套的抽屉。
- 我们假设原先左手手套和右手手套的数量相等——假设总量相等。
- 如果这样的假设(虚无假设)是正确的,我们已计算出了取出五只右手手套的概率。

9.3 概率要低到什么程度

我们发现在虚无假设下得到的概率是非常低——事实上是0.03。这看似很低,但p值多低才能成为令人信服的证据。

这有一个所有研究都用的值,如果他们计算在H_0下结果发生概率的p值低于这个值,他们就拒绝H_0并假定有真正差异存在——支持H_1,即有效应存在。这个值是什么呢?很多人都对p赋值,但它并非是任意的,让我们来做以下的练习:

猜婴儿的性别

假设一个朋友说他能通过在母亲子宫上方摆动的石钟来预测未出生婴儿的性别。让我们来假设她能准确地猜出你孩子的性别。你会受其影响吗?你可能会做出“惊讶”反应,或至少认为“这很有趣,可能有一些什么东西在里面”。回到刚才的话题,如果她仅仅是猜测,她有一半的机会是正确的。然而,大部分人开始想,如果她能够成功地预测准确2~3个或更多朋友孩子的性别,她就确实有一些能力。假设我们让其在一个严格控制的科学条件下猜测10个孩子的性别。你认为她能预测正确多少才能说不是凭猜测和运气呢?比如,10个中准确预测7个会让你信服吗?或者你希望更多或更少呢?

我设想在这种情况下,大多数学生认为10个中有9个或10个正确才能使他们信服,同样我们看到,社会科学家希望结果的完美巧合是为了表明他们的结果是“显著的”。如果她仅有8个是正确的,观众就开始犹豫不决了,但也有可能还有1/3的人还是相信的。一些充满幻想的人认为7个正确,甚至6个正确就足以使他们相

① 在一些高水平的书上也有方法的错误。若你接受这些论述,认为每个都是正确的,那么学习本章请把它当做向私人教师来阅读。

信她拥有预测的能力，但也有一些玩世不恭的人认为即使预测 10 个都完全正确也不能提供足够的证据。

9.4 常规的 0.05 显著性水平

社会科学家和其他大部分涉及统计显著性检验的人接受 0.05 是拒绝虚无假设的合理的限制水平。也就是说：

在假设虚无假设成立时，如果得到结果的概率小于或等于 0.05，就拒绝虚无假设。

现在你会认为一般个体对这种合理的水平有个直观的理解。我的大部分学生观众十分相信她猜中了 10 个中的 9 个或 9 个以上就算成功了。它发生的概率是 0.011。你可以在图 9.2 里看到这个值，每个圆柱形都代表从 10 个中猜中数量的概率，猜中的数量显示在圆柱形的底部。当然，这个概率是在这样的虚无假设的情况下发生，即对性别的预测是完全随机的。一半不到的观众可能很犹豫地接受如果她仅仅能猜中 10 个中的 8 个或 8 个以上，她就是有能力预测婴儿的性别的。有趣的是，这里的概率是 0.055，仅超 0.05 水平的显著性。①

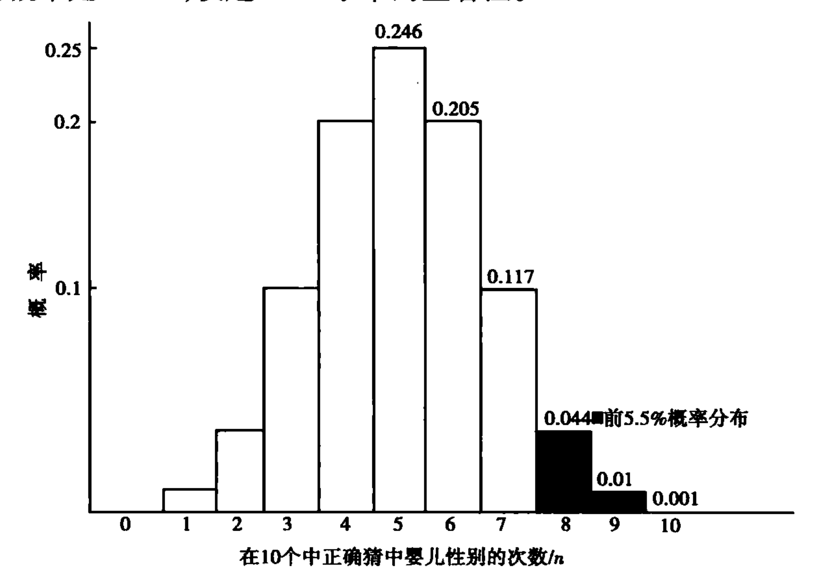


图 9.2 猜中婴儿性别的概率

9.5 临界值

我们看到如果我们从 10 个中猜中 8 个或以上，就不会拒绝虚无假设。如果我

---

① 因为我们把猜中 8 个、9 个、10 个的概率相加，所以，这里概率是 0.044 + 0.011 + 0.001，并不是 0.044。我们不会预测她恰好猜中 8 个，但知道她会做得更好。我们必须计算她猜中 8 个或猜中更多一些的概率。9 个正确的概率值是 0.01 + 0.001 = 0.011。

们猜中 9 个,概率就降到 0.01,就在可接受的 0.05 水平之下了。猜中 9 个或以上的概率值就叫**临界值**(critical values),因为这是我们拒绝虚无假设时必须达到(或更好)的数值。

练 习

1. 一个研究者报告 $p=0.049$ 差异是显著的。她的对手会对此说什么或做什么呢?
2. 另一个研究者决定公开研究结果:在虚无假设下它们发生的概率都小于 0.1。他的同事和对手会对这项研究说什么呢?
3. 一个研究者以苏格兰和英格兰 8 岁儿童为样本进行阅读测试。她想找出两组群体的差异。在这个统计检验里的虚无假设是什么?
4. 你有两个装有成千上万不同长度铅笔的盒子。你被告知虚无假设是正确的——两个盒子里铅笔的平均长度是相等的。你从一个盒子里取出 30 支笔作为样本,从另一个盒子里也取出 30 支笔作为样本,如果你比较两个样本的平均数,你希望找出什么呢?如果你这样操作 40 次,你认为会发生什么?

答 案

1. 他们会说:"那仅仅是制造出的。"他们可能会重复她的研究。如果得到一个不显著的结果。他们可能会认为她的结果仅是侥幸的成功,见 9.8 节的 **I 型错误**(Type I ereor)。
2. 他没有获得可靠性(credibility)。0.05 是常规的水平。用 0.1 水平就使获得侥幸结果的机会更多、更严重。他更有可能犯了 I 型错误。
3. 这个测验里 8 岁苏格兰儿童的总体平均数与 8 岁英格兰儿童的总体平均数相等。
4. 你可能希望(并预测)两样本的平均数接近。它们之间的差异很小。它们之间的一些差异是由于样本误差导致的。如果你反复这样做很多次,你会认为平均数之间的差异很小。另外布朗尼(Brownie)指出如果你注意并说出每 20 次中可能会有 1 次,你会得到"显著性"差异。

关键术语

备择假设(H_1)(alternative hypothesis, H_1)	研究预测(research prediction)
临界值(critical value)	显著性(significance)
实验假设(experimental hypothesis)	显著性水平(significance level)
虚无假设(H_0)(null hypothesis, H_0)	

9.6 以真正的心理学研究结果为例

到目前为止,我都只是通过运用硬币,手套和猜婴儿性别来介绍假设检验的思想!认真的学生可能会想了解,什么时候才能与典型的心理学试验相联系呢?回答是"立刻",我想通过一个准确的例子来介绍假设检验。

不可靠螺丝的例子

假设你在一个生产螺丝的工厂工作,任务是照看生产线,找出螺丝,装入能盛 500 000 个螺丝的桶里。一天你的监管走到你的身边说:"我们估计切割机掉到线外

了，我们怀疑那边环绕黄线的桶里的螺丝没有被精确地切割，你能替我们检查下吗?”首先，你会想“我的周末就这样没有了”，因为你想象着坐下来检查桶里的每个螺丝。但是老员工会告诉你怎么做——从装有可能非标准螺丝的桶里找个相当好的样本，再从其他标准的桶里找个相同规格的样本，用假设检验来比较两者的差异。

如图9.3描述的那样，我们用一个“相当好的”样本是什么意思？这就涉及我们在第2章讨论的样本问题。为了能公正地代表整个桶里的螺丝，我们要随机地抽取我们的样本，而不是简单地从表面抽取，表面可能是小螺丝集中的地方。我们说相当规模的样本意思可能是20或30个，绝不仅仅是1个或2个。我们抽取出这些样本，算出每组的平均数并进行大小比较。

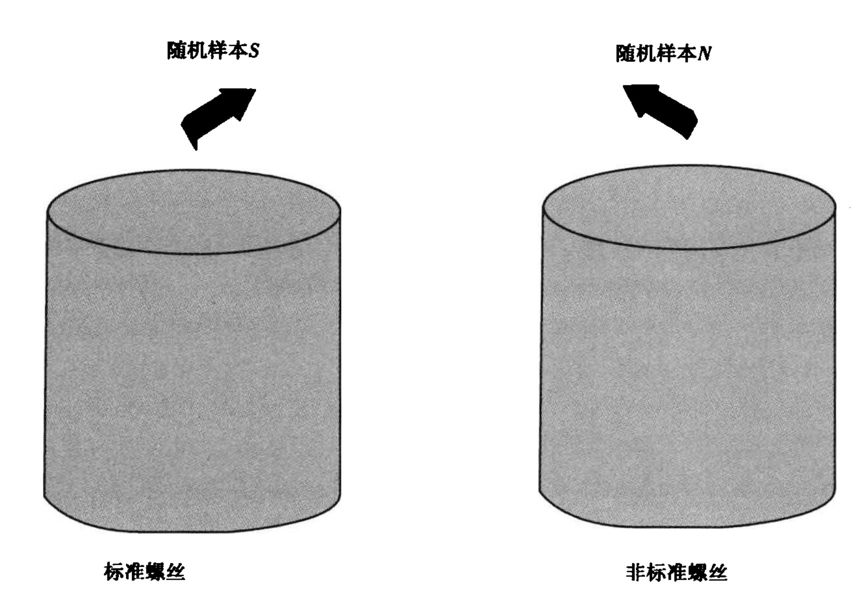


图9.3 标准和可能非标准螺丝的总体

如果仅仅因为固定的样本误差(sampling error)——常常是极小的差异，其中一个平均数会变大。但是，我们要寻找的差异要比我们希望的从相同总体中任意取出的两个样本之间的差异要大得多。

在这个例子里虚无假设是什么？如果两个桶里装的是以同样的标准制造出来的螺丝，两个桶里的平均数应该相同。桶是“总体”，我们从这里选出样本。我们检验的虚无假设是，标准桶里螺丝总体的平均数和可能非标准桶里螺丝总体的平均数相等。我们的备择假设是两总体的平均数不等。

重回以真正的心理学研究结果为例

通过螺丝的例子我们开始得到真正的心理学研究结果。让我们重新回顾本章初始部分表9.1列出的数据。认真学生所在组的数据呈现在表b的部分。这里你要看的两个样本数据同以上螺丝例子的两个样本一样，不同的只是这里的螺丝总数是有限的。我们知道每个桶里有许多螺丝，但是它们的数量是固定的。在社会和心理学的统计检验上，很少有总体数量是真正有限定的例子。我们讲的“总体”，在统计上是指“所有以这种方式存在的数据”。

用心理学的研究结果概括

虚无假设 H_0:"观察"条件下所用时间的平均数 = "单独"条件下所用时间的平均数

备择假设 H_1:"观察"条件下所用时间的平均数 ≠ "单独"条件下所用时间的平均数

这两个假设用于每个样本数据的总体。

从 H_0 我们能预测我们的研究中,两样本平均数之间仅有很小的差异。

我们可以计算在虚无假设下我们得到样本平均数之间存在差异的结论的概率的大小。我们将在第 11 章学习怎么计算。这里假设我们已能做到这一点。

如果概率(p 值)小于等于 0.05,我们就拒绝虚无假设,推断有显著差异,支持备择假设——有观察者时则增加了犯错误的时间。

注意:当我们正确地知道我们的数据应用哪种统计分析时,才能建立虚无假设。典型的虚无假设是:

- 两总体的平均数相等。
- 两总体的中位数相等。
- 在总体里变量 A 和变量 B 的相关为零。

9.7 对 0.05 显著性水平的解释

0.05 显著性水平(significance level)有两个要点。第一个是实践性。如果一个结果在 0.05 水平上显著(有时标记为 **5% 水平**),这通常意味着其他研究者也接受这个结果,我们可以在学术期刊上发表这个结果——"公开发表"。我们假设的效应在这里是显著的——例如组间的差异和相关(见 12 章)。如果未达到 0.05,会发生什么? 当然我们不是武断地"得到结果"。同样我们看到,0.05 水平仅是令人信服的水平,这并不是效应的"真实性"所在。如果一项研究勉强不能达到显著性,可能会用更加严格的统计方法来重复研究,或有更多的参与者来做,直到达到显著性,并提交发表为止。然而,在一些文章里研究的目的并不是呈现差异性。我们可能想说明两组之间并不存在差异,正如世俗或社会期望所希望的那样(例如,试图证明性别差异)。我们可能想表明一些人表面上的论证并未起作用。例如,我们的理论可能要求性格外向的人与内向的人在某些变量上不存在差异,但他们确实在其他一些变量上有差异。

这就涉及第二个要点。0.05 的限定并不是效应存在的保证。结果可能仅是一个侥幸。我们怎么做才能防止侥幸的发生呢?

9.8 Ⅰ型错误和Ⅱ型错误

因为我们设置的显著性水平是 0.05,所以每当我们做 20 次随机测验,平均这 20 次中应该有 1 次是显著的。如果我们分别从数量相等的两个盒子里随机抽取 30 张有奖活动的票,并检验是否一个的总数比另一个的多。我们应该在 20 次检验中

有1次会得到“显著”差异。定义显著性水平的另一种方法，就是看它做了什么。大部分研究者不做随机检测，因为他们已在理论中论证，他们通常希望在样本之间存在差异，但是，这常常没有真正的效应，除非研究者偶然得到一个“显著”结果。我们将这种情况称为Ⅰ**型错误**（type Ⅰ error）。当然我们不知道什么时候会发生。我们仅从重复的研究中发现。如果我们要得到一个效应，在紧接着仔细完成的十项研究中却并没有得到，那我们就可能认为结果仅是Ⅰ型错误。

在表9.1的a部分里出现的马虎学生组的结果是怎样的呢？可能是观察者的效应（事实上常常是这样），他们的设计是如此的马虎以致没能显现出来。如果是这种情况，就称为Ⅱ**型错误**（type Ⅱ error）——有效应但这个结果没有达到显著性。Ⅰ型错误和Ⅱ型错误在表9.2中呈现出来了。

表9.2　Ⅰ型错误和Ⅱ型错误

	保留 H_0	拒绝 H_0
H_0为真	正确决断	Ⅰ型错误
H_0为假	Ⅱ型错误	正确决断

因此达到显著性并不是所有事件的“证明”。我确信我已说过你应当从心理学写作中删除“证明”一词。一个显著性结果提出了足以支持我们理论的证据；我们需要其他的证据，其他不成功的研究来反驳我们的结果，更加牢固地支持我们的理论。另一方面，未能达到显著性并不是效应不存在的“证明”。它仅仅表示我们没有增加任何更多的证据或我们反驳了其他研究者的结论。在所有这些结果中，研究者的工作是尝试解释为什么研究结果支持或没有支持该理论。

9.9　其他水平的显著性

我想我会重复在显著性检验里计算 p 值的意思。它是在确定没有任何事情发生的情况下得到你所看到的结果的概率（比如，差异如此之大）。如果的确是没有人在夜里戳破你的轮胎，就可以把它想象成在一周里获得三个光滑轮胎的概率。如果我们用0.05的显著性水平，即使虚无假设成立，没有效应存在，那么100次里有5次，你就会得到一个“显著”结果。例如，100次里你有5次使5个或更多相同颜色的纸牌出现在一副洗好的纸牌的最上面的概率；100次投掷10个硬币，你可能超过5次得到8个正面向上的概率。

使用 $p \leqslant 0.01$（1%水平）

有时候研究者对他们这样一个事实感到不满，即对研究结果发生的概率只有5%这个事实。有时他们想让犯Ⅰ型错误的概率小些。因此，他们所做的是，如果结果在虚无假设下的概率小于0.01，他们就拒绝虚无假设。这就意味着犯Ⅱ型错误的概率增加。换句话说，例如，可能房间里的温度并不影响记忆。但是如果显著性水平设置得低至0.01，即使温度真正影响记忆，结果也会被认为不足以拒绝虚无假设。$p \leqslant 0.01$ 是一个“严格”的水平。当一个研究者发表某一领域突破性的报告，宣

称他颠覆了其他研究者的研究时,我们可能选用这一显著性水平。如果你要挑战别人已经发现的东西,那么你需要一个真正坚实的基础去支持你的结论。选用一个低的显著性水平的另一种情况是,你只有一次机会获取数据。在一个现场研究中,我们需要考察在学校中实施"反对欺负弱小者"计划的效果,在这里可能只有在一种情况下才能对不同学校作出公正的比较,实施和不实施该计划。这个事例告诉我们,既然其他研究者不能复制我们的研究,就应该公正地确保任何一个被主张的效应是真实的。在我们的研究结果可能被应用在真实环境下对人进行治疗(例如,药物治疗)的情况下,我们也需要一个严格的显著性水平。

使用$p \leq 0.1$(10%水平)

如果结果显著性仅为$p \leq 0.1$,很少的期刊会公开发表这样的研究结果。但是,在准备性的工作上,一些研究者可能尝试不同的方法来检验一个假设,p值比0.1小的设计可能会被留下,然后再修改,这是为整个测试作准备,希望最终得到一个在0.05水平上显著的结果。表9.3概括了不同显著性水平对应的特性。

表9.3 不同显著性水平的使用

显著性水平	解 释	在虚无假设成立情况下,犯Ⅰ型错误的机会
$p \leq 0.05$	研究常用的水平。	5%
$p \leq 0.01$	"严格"的水平;很难得到显著的结果。用于进一步确定结果效应的真实性。	1%
$p \leq 0.1$	针对学术上对真正效应的接受而言太高了;可用做实验性工作的指导。	10%

9.10 术语"显著性"的使用

如果一个呈现的结果并不显著,请不要用"不显著的结果"来描述,这就混淆了假设检验中统计技术语言的"显著"和日常用语的"显著"。许多事件在日常生活中是"显著的",但在统计上并不是"显著的"。心理学研究的许多结果是显著的并不等于大多数计划中事情都是这样的。在"一个非显著的结果"里,我们使用的术语是"非显著的结果"。研究者常常说:"控制和实验之间的差异意味着不能达到显著"。一个非显著的结果在非统计意义上常常是显著的。未能呈现其他研究者的陈述是显著的事件,除非"未能呈现"是带有"非显著的统计结果"的证据。

更多的术语

"拒绝虚无假设"

"保留虚无假设"

"暂时接受备择假设……"

"找到了支持备择假设的证据"

我们从不说备择假设是"被证明了的",我们在心理学里不简单地去"证明"什么。事实上,"证明"主要用于数学范围内,同所有其他科学一样,在心理学上,我们寻找假设的支持,或怀疑先前的支持,或我们没有找到支持。

9.11 单侧和双侧检验

从图9.2中你会看到我们仅仅计算了猜中婴儿性别8个或更多的概率。概率是图右阴影部分的面积,这被称为单侧检验(one-tailed test),因为就像你从图中看到的,我们仅仅用了概率分布的一侧。我们应该知道找出8个或更多正确和8个或更多错误的概率,在做这个时,我希望你明白,我们必须把左边的概率和右边的概率相加。我们得到的概率会是原来的双倍,从0.055到0.11。但是,这样做显得很多此一举,我们只关心猜对的概率程度。

但是当测量心理学研究的其他问题时,比如当"观察者的出现是促进了还是阻碍了被试的表现"时,我们就要用**双侧检验**(two-tailed test)。这种校验的类型直接与假设的类型有关——定向的或非定向的。在**定向的假设**(directional hypothesis)中,我们预测结果将朝哪个方向发展——例如,"有观察者存在的情况提高了成绩"。在**非定向的假设**(non-directional hypothesis)中,我们不能作出定向的评价——"咖啡因会影响人的反应"。检验和假设的关系如下:

假　设	检　验	显著性水平
定向的	允许单侧	双侧检验的一半
非定向的	必须双测	单侧检验的双倍

当你在以下章节进行推断检验描述时,当你在附录表中寻找概率值时,你常常需要决断是用单侧还是双侧检验。基本上来说,用双侧检验更为保险些。就像你刚看到的,概率在双侧检验中是双倍的,这就意味着如果你使用单侧检验的结果是显著的,在你使用双侧检验时,结果可能就不显著了。单侧检验的显著性是$p \leqslant 0.05$,双侧检验相应的p值是0.1,这并不是一个显著的结果。

用单侧和双侧检验回答测试的问题

但是,我们卷入了统计学和心理学的研究者之间令人难过的争论之中。作为心理学的一般法则,研究者几乎都使用双侧检验。原因很复杂,但我们能从库里坎2004年著作中的第11章里找到原因(Coolican,2004)。

在准备本书时,在四种高水准的题目要求中,至少有两种仍然要求学生回答关于单侧和双侧检验的问题。他们希望学生能说明当假设是定向的时候必须要用单侧检验。

在一个测试里,假如你被一个研究者告知,他预测狗在训练过之后看到水平条纹毛衣时,吼叫得比看到平针毛衣时更厉害,问题就出来了。"研究者是用单侧检验还是双侧检验",我恐怕你对这个问题的回答是"单侧检验,因为研究者对结果发展偏向作出了预测"。如果你感到自信,你常常还会接着说:"但是,事实上几乎所有的研究者都用双侧检验。"这仅仅是测试的编写者没有真正与研究的真实世界联

系。更多的真实环境是我们能预测到的，比如，音乐能影响我们的记忆，我们可能惊讶地发现记忆能力提高了。我们想知道得更多，双侧检验的显著性使我们可以说，尽管得到结果不是我们希望的，但它是有效应的。

练 习

1. 说出下面的假设检验是用单侧检验还是用双侧检验。
   (a) 糖尿病患者对健康的担忧比非糖尿病患者的多；
   (b) 外向的人和内向的人对自尊的定位不同；
   (c) 焦虑与声望呈负相关；
   (d) 与事故后果轻微相比，当事故严重时，事故的责任会更多地归结于驾驶者；
   (e) 人的慷慨与性格相关。
2. 一个研究者指出：对已发表的与他们的观点相反的言论关注更多的人，会比那些关注更少的人更容易改变他们的态度。这个预测与先前的研究认知失调(cognitive dissonance)的结果背道而驰。
   (a) 对于她使用什么样的显著性水平是合适的？
   (b) 用这个水平是不是与用通常的 0.05 水平犯 I 型错误差不多？
3. 我必须给你这些问题(它们会在测试中出现)：
   (a) 在假设检验里 p 是什么？
   (b) “用 $p \leq 0.01$ 表示差异是显著的”是什么意思？

答 案

1. (a) 可能是单侧检验(定向假设)
   (b) 一定是双侧检验(非定向假设)
   (c) 可能是单侧检验(定向假设)
   (d) 可能是单侧检验(定向假设)
   (e) 一定是双侧检验(非定向假设)
2. (a) $p \leq 0.01$
   (b) 极小可能
3. (a) 如果虚无假设成立，p 是这个结果出现的概率
   (b) 在虚无假设下，结果出现的概率小于 0.01

关键术语

5% 水平(5% level)	显著差异(significant difference)
定向假设(directional hypothesis)	双侧检验(two-tailed test)
非定向假设(non-directional hypothesis)	I 型错误(type I error)
单侧检验(one-tailed test)	II 型错误(type II error)

10

显著性检验的选择

本章内容

- ❑ 本章只有一个中心主题——帮助你选择合适的显著性检验。
- ❑ 本书中各类显著性检验本章都涉及了，但其更详细的阐述请参见第 11 章或第 12 章相关部分。
- ❑ 选择每一种检验都基于三个决定因素——要检验的数据关系的类型、变量的测量水平以及设计的类型；本章对三个因素作了阐释并在图 10.1 中表示出来。
- ❑ 阐释了参数检验和非参数检验之间的区别。
- ❑ 练习部分为你提供了决策实战的机会。

10.1 数据检验

你已收集了数据，并想知道数据是否支持你的假设。要能得到数据支持研究假设的结论，你必须要证明，如果真的没有效应存在，你从数据中得到这个结果（差异或相关，a difference or a correlation）的可能性极小（$p \leqslant 0.05$，这一点在上一章解释过）[①]。你需要选用什么样的检验只依赖于三个决定因素：

1. 要检验的数据关系的类型——差异还是相关。
2. 数据的测量水平。
3. 研究设计是相关还是独立（related or unrelated）。

以上三个决定因素依次明确后，就可以为你的数据选择合适的检验了。图10.1涵盖了第11章、第12章涉及的所有检验，并给出了如何考虑上述三个决定因素的步骤。不同的检验，数据的整理格式不同，这一点可以参考图10.2。请注意，一旦你选择了这里提供的合适的检验，你就可以参考表11.2，它给出了进行**任何**显著性检验的一般程序。

差异还是相关

如果是进行差异检验的话，需要有两列分数。目的是检验一列分数比另一列分数高。有一种特殊情况是，数据是**分类**（categories）的，你想检验两组在两项或多项分类上是否存在差异。比如，观察男性和女性，并记录他们是否带书。又比如，观察汽车不适当占用受损车位的情况，并记录每辆车的类型（例如，是豪华、中档还是便宜的？）。不论何时，只要你用这样的方式进行记录，观察频次，你可能要用到卡方检验（Chi-square test）。这样你检验的是每种分类的频次之间的差异。若你用到这种检验，这里不多讲了。

相关检验是回答你提出的诸如“身高和自信之间有关系吗——高个子是否更自信？”或“外向的人是否倾向于晚睡——内外向与就寝时间是否有关系？”之类的问题。相关分析寻求的是两个变量之间的关系。只要每个被试在两个变量上都各有一个分数，相关分析就可以评估两列分数共同上升或下降的程度。相关分析就其本身来说是一个重要主题，这将在第12章里单独处理。

数据的测量水平

如果每个被试的得分都是测量分数（scores），这样的数据就是等距水平。等距水平的检验（interval-level tests）（以下称之为参数检验）要求数据满足某种假定条件，而且如果不用计算机的话，*t* **检验**计算起来相当棘手。因此，在这种情况下，完全可以优先选用顺序水平的检验（ordinal-level test）（“非参数的检验”）而不用等距水平的检验：如果是差异检验，可以优先用**曼-惠特尼** *U* **检验**（Mann-Whitney *U*）或**威尔科克逊检验**（Wilcoxon's *T*）而不用 *t* 检验；如果是相关分析，则用**斯皮尔曼**（Spearman）

---

① 反之，小概率事件出现了，我们就有95%的把握拒绝没有效应存在这一虚无假设，从而支持研究假设。——译者注

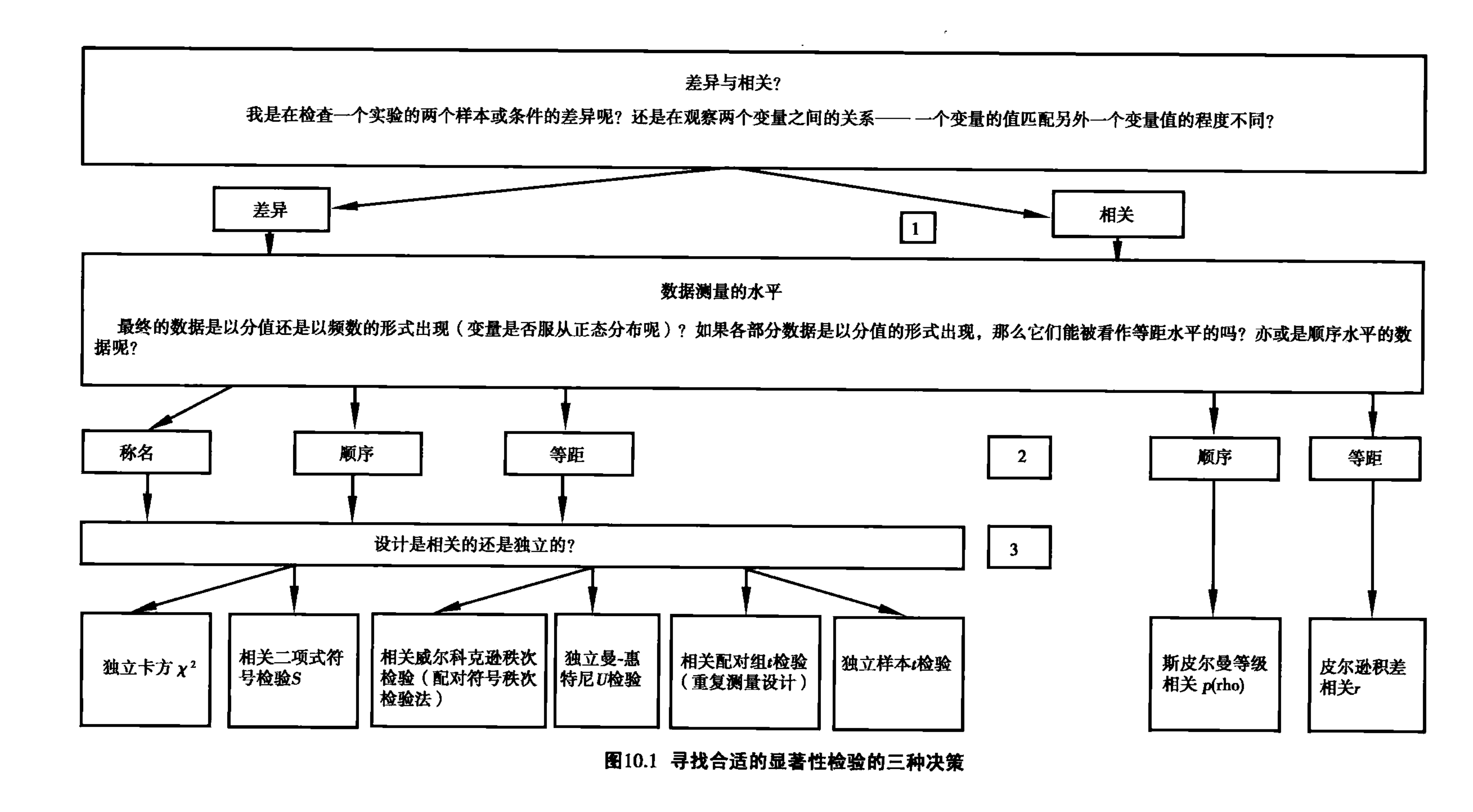


图10.1 寻找合适的显著性检验的三种决策

检验两种条件之间的差异			
独立分组		同一组（或配对）	
样本A	样本B		
条件1	条件2	条件1	条件2
3	6	12	23
5	7	13	12
6	5	15	24
⋮	⋮	⋮	⋮
独立t检验 **曼-惠特尼检验**		**相关t检验** **威尔科克逊配对样本符号检验**	

相 关	
同一组人（或配对）	
测量1	测量2
12	23
13	12
15	24
⋮	⋮
皮尔逊积差相关 **斯皮尔曼等级相关**	

关联或差异			
	汽车类型		
	豪华	中档	便宜
泊车	23	19	12
不泊车	10	24	21
卡方检验			

图 10.2 各种假设检验的数据整理格式

等级相关而非**皮尔逊**(Pearson)**积差相关**。非参数检验不如参数检验那么有力(powerful)①,因此,有可能独立样本 t 检验结果显著,而曼-惠特尼检验不显著。不过,如果研究设计良好,你应该还是可以相信你得到的结果。而且,96%的情况下,两种检验的结果相同。如果非参数检验的结果接近显著,你不妨用 t 检验再做一下。

对于每个被试都有一对分数的情况进行差异检验时,除了上述将数据降为顺序水平的方法外,也可以把数据降为称名(或曰名义)水平进行**符号检验**(sign test),方法是简单记录每个被试由一种条件到另一种条件时情况是好转、变坏还是维持原状。然而,这是不明智的,因为这样做会丧失掉很多信息,其检验力大大弱于威尔科克逊法或相关 t 检验。不过,符号检验可用于如下情景:你所知道的关于每个被试的全部信息仅限于他们会给出一个正号、负号还是零——比如,你询问被试他们对

① 所谓有力,是指统计检验探测差异的能力,一个有力的检验比一个不大有力的检验对数据里的差异更敏感。——译者注

狩猎是同意,还是不同意,抑或不发表意见①。

一些技巧

- ❑ 如果对每个被试都进行计分——就是说,由于他们做了或选了 *X* 而非 *Y*,而被计入某一"格 "②——那么,数据是称名数据,用卡方检验是合适的。
- ❑ 如果因变量(dependent variable)(用以测量被试)是一个标准测验,则数据可能是等距水平。
- ❑ 如果你让被试用主观判断(subjective judgement)进行评定(比如,"用 1 至 10 的等级标明你在多大程度上支持捕猎狐狸?"或"……某人有多大的吸引力")。那么,这种情况下,不要将数据视为纯粹的等距数据,而将其视为等级排列数据进行检验更保险。

研究设计是相关还是独立

由于**所有的**相关分析都用相关设计,因此当你做的不是相关分析的时候才需要考虑研究设计是相关还是独立。若每个被试在自变量(the independent variable)的**两个**水平下都作了测试,这就是相关设计。若被试是配对(paired)的,每对被试中的一个在自变量的一种水平下测试,另一个在自变量的另一种水平下测试,这样的设计也是相关设计。若非如此,设计就是独立的——独立设计通常意味着将被试分为两组,一组在自变量的一种水平下进行测试,另一组在自变量的另一种水平下进行测试。独立设计还有一种情况,即用因变量(比如摩托驾驶技能和阅读能力)测试两类不同的人(典型的是男性和女性)。就测试来说,男性组中的每一位男性显然独立于女性组中的每一位女性(即使你很可能碰巧测试到了兄妹!)

10.2 什么是参数检验

之前多次提到参数检验,现在解释一下什么是参数检验以及参数检验为何优于其他类似的检验。参数检验是通过**样本统计量**(sample statistics)来估计**总体参数**(population parameters)。由于参数检验依赖于对分数的潜在分布作出估计,所以其"**依赖分布的检验**"(distribution-dependent tests)之名可能更为人所知。一般来说,我们更愿意用参数检验,因为参数检验通常比相应的非参数检验更有检验力(但非参数检验与参数检验在 96% 的情况下检测力相当)。这意味着,对于同样的数据,参数检验更可能达到显著性差异。相对于非参数检验,它对数据更加灵敏一点;它利用了更多可利用的信息。从长远看,我们更可能发表同行接受的支持效应存在的论文。非参数检验则不依赖于对潜在分布作出假定。

虽然参数检验更灵敏,更易得到显著性结果,然而,这是有代价的。由于参数检

---

① 将被试的回答按级别编码,比如,将"不同意""不发表意见""同意" 依次编码为 1、2、3,然后用每个被试第二次测量的数字减去第一次测量的数字,用" + "" - ""0"记录,进行符号检验。——译者注

② 可以参见图 10.2 底部右端的例子进行理解。比如,一辆豪华汽车没有在受损车位泊车,则它被计入数字为"10"的"格" 中。——译者注

验是对总体参数作出估计,如果数据可能使这种估计不可靠时,我们就要慎用这种检验。因此,对于本书中用到的参数检验,要求数据集(dataset)满足以下假定条件:

1. 数据的测量水平至少是等距的。
2. 样本数据应该来自**潜在的正态分布总体**(an underlying normal distribution)。
3. 两样本的方差(variance)差异不显著——此假定即为熟知的"**方差齐性**"(homogeneity of variance)。"齐性"意味着"相同类型的",因此,这意味着我们假定方差是相同的。

- ❑ **假定1——等距数据**:注意,我们之前讲过,收集的数据并不总是顺序水平的。当我们认为数据不是真正的等距数据时,就要给数据排序。当收集的数据是人为估计(huaman estimates)的时候,这样做就是明智的。
- ❑ **假定2——正态分布**:抽取样本的潜在总体必须假定为正态的。然而,我们得不到总体,也很少知道总体是否正态。不过,如果运用了标准化的心理量表(a standardized psychological scale)来测试变量,我们就知道总体在此变量上的分布是正态的。用标准化量表测量的结果能形成一个正态分布(a normal distribution)。一般可通过大体上观察偏态(skew)来检验正态性。也许你不需要计算偏态,但可以通过简单观察数据的直方图(histogram)中明显偏离常态的点的情况来进行判断。
- ❑ **假定3——方差齐性**:一般情况下可以不理会此假定,但下面的情形除外:独立样本(被试的分数分为两组),两组人数很不均等。这种情形会导致方差不齐,一个简便的解决方法是确保每种条件下被试人数大致相当——不必精确相等。若你确实想检查方差是否齐性,一个法则是:如果一组的方差超出另一组四倍多,那就有问题了。记住,方差是标准差(the standard deviation)的平方。

可能你需要验证选用的检验是不是正确合理。对于参数检验,要检查数据是否符合上述所说的假定。若你不想检查,而且数据情况允许的话,可以选用非参数检验。前面说过,非参数检验通常更易于计算。除非规定用参数检验,何不考虑采用非参数检验呢![①]

练 习

为了帮助你理解选择显著性检验的程序,同时也看看你是否已学会选择显著性检验,这里提供一些心理实践工作中最普通的假设检验以供练习。看你能否决定选择哪一种合适的检验,表10.1提供了答案供你参考。

1. 提供一列单词,要求一组被试复述并记住每一个单词;另一组被试对每一个单词进行生动的想象。两分钟后要求两组被试回忆单词。想象能否提高正确回忆单词的数量?
2. 测试被试的外向程度,然后让其在"公共场合演讲时感觉到的自信程度"的量表上给予1至10级的评定。我们预计,更外向的人会更自信。

---

① 本书作者似乎更倾向于采用非参数检验,但从国内外研究文献看,还是采用参数检验的较多。——译者注

3. 学生从当地报纸上搜集了征婚广告来检验这样的假设:女性比男性更可能提及长期关系和安全感概念。他们关注的是征婚者的性别和是否提及安全等。
4. 一个研究者指出,人数更多的班级是导致数学成绩更少得 A 的一个原因。她运用全国性的 A 级结果,然后对每一个足够大的学校,记录了数学班的平均人数和平均 A 等数学分数。
5. 要求被试在有观众在场和独自一人两种条件下完成分类任务,并在很短的时间内测试成绩。预计有观众在场条件下的归类成绩较差。
6. 学生记录汽车在黄色交通信号灯前是否停下,同时记录每辆轿车的价格范围:便宜、中档或昂贵。他们预测,汽车的价格越高,司机越可能停下车。
7. 学生观察男性和女性是否在梯子下走。他们将要检验这样一个假设:一种性别比另一性别更迷信。
8. 要求男性和女性估计他们自己的 IQ 和他们父母的 IQ,同时对他们自己和父母实际的 IQ 进行了测试。对以下假设分别选用什么显著性检验?
   (a)男性对自己 IQ 的估计高于女性对自己 IQ 的估计。
   (b)被试对父亲 IQ 的估计高于对母亲 IQ 的估计。
   (c)实际测试到的父亲的 IQ 与实际测试到的母亲的 IQ 有差异。
   (d)母亲的 IQ 越高,其儿子的 IQ 也越高。
9. 在两种独立的条件下评估一组被试的抑郁程度。没有原始分数,我们仅知道每位被试的情形是变好、更坏还是维持原状。哪种检验能告诉我们,被试在一种条件下是否比另一种条件下有显著性的好转?

答　案

参见表 10.1。

关键术语

依赖分布的检验(distribution-dependent tests)
非参数检验(non-parametric tests)
方差齐性(homogeneity of variance)
参数检验(parametric tests)

表 10.1 选择合适检验问题的答案(黑体字表示优先选择这种检验,因其更有力)

问题	决定因素 1 差异还是相关	决定因素 2 数据水平	决定因素 3 相关还是独立设计	可用的检验
1	差异	等距(单词数量)	独立——每种条件下是不同的被试	**独立 t 检验** 曼-惠特尼检验
2	相关	顺序——测量是人为估计(见 p. 108)	相关——每位被试有两个分数。所有的相关分析都是相关设计	斯皮尔曼相关
3	差异(男女之间)	称名——男性/女性;每一个测量要么提到安全感要么没提到	独立——男性对女性	卡方检验
4	相关	等距	相关——所有的相关分析都是相关设计	**皮尔逊相关** 斯皮尔曼相关
5	差异	等距	相关——每个被试两种条件下都作了测试	**相关 t 检验** 威尔科克逊符号检验
6	差异——不同的汽车类型间	称名——汽车类型与是否停下	独立——不同的汽车类型	卡方检验
7	差异	称名——在梯子下走或不走;男性/女性	独立——男性对女性	卡方检验
8a	差异	顺序——人为估计	独立——男性对女性	曼-惠特尼检验
8b	差异	顺序——人为估计	相关——同一组人两列分数	威尔科克逊检验
8c	差异	等距——测试的 IQ	相关——父母是配对的!	相关 t 检验
8d	相关	等距——测试的 IQ	相关——母亲和儿子。所有的相关分析都是相关设计	皮尔逊相关
9	差异	称名——没有分数,只有分类	相关——原本是每组被试两列分数	符号检验

11

差异检验

本章内容

- 本章介绍了为寻找两组数据间差异时所需的所有显著性检验。同时也介绍了适用于一个或两个变量的频数分布表的检验。
- 我们首先大致介绍一下进行任何一种显著性检验的步骤。
- 本章包含了以下检验：
  ——相关 t 检验
  ——非相关性 t 检验
  ——威尔科克逊配对符号秩次检验
  ——曼-惠特尼检验
  ——χ^2 检验
  ——二项符号检验
- 列出了每一种检验的使用条件和要点解释。这些条件通常是对使用某种特定检验的合理性作出解释时所必需的。
- 以一个处理过的案例为例，给出了对每一种检验的简短解释以及它的基本原理，同时也给出了报告每一种检验结果的规范方法的示例。

回顾上一章的开始部分，你会看到我们提出了这样一个问题：在实验中发现的组间差异能否被看做是一个"真的"差异。现在我们知道可以将一个差异看做是潜在的真效应(underlying real effect)的临时性证据。如果没有任何效应，那么这个差异发生的概率低于0.05——更规范地说法是在虚无假设的条件下差异发生的概率低于0.05。在接下来的两章中，我们将探讨为了计算出那个概率你可能需要的各类检验。

我们使用的检验被称为统计**显著性检验**(significance tests)或**推断检验**(inferential testes)，使用这些检验时，我们仅仅是从样本中已经得出的结论去推论存在于总体中的效应。当我们进行某一推断检验时，需要经过以下几个步骤：

1. 搜集数据并决定它适用于什么检验(参见第10章)。
2. 进行检验。
3. 计算检验统计量(test statistic)的值。
4. 决定采用单侧检验还是双侧检验。
5. 在虚无假设下获得该检验统计量的概率。
6.

如果概率小于或等于0.05就拒绝零假设(H_0)，说明存在显著性差异(或相关)。	如果概率大于0.05就保留零假设(H_0)，说明差异(或相关)不显著。

第10章试图提供选择一个合适的检验所需的信息。然而，我们在这里给出的每种检验的例子也将帮助你决定你的数据适合于哪种检验形式。我们将给出进行每一种检验所需的步骤，而没有必要给出为什么要这样做的深层次的统计学解释。我们认为在这个阶段，你只是被要求去分析数据而不需要去准确地理解每一种检验是如何进行以及为什么是这样的。关键是在所有检验结束后，需要报告你的检验统计量，你的检验是单侧的还是双侧的，与你的检验统计量相关联的概率以及这一概率是否使你有足够的信心拒绝虚无假设(H_0)，这些将在关于报告写作的第14章中得到补充。第14章也将告诉你在结果部分我们还想要看到别的东西，但在这章我们只是告诉你如何报告检验结果。

11.1 差异的参数检验

相关样本 t 检验

首先让我们来看这样一种检验，这种检验是因计算统计量 t 而得名的一组检验中的一个。这个检验是威廉·戈塞特(William Gossett)发明的。戈塞特为吉尼斯(Guinness)工作感到十分不便，因为在那个时代，吉尼斯不允许戈塞特以自己的名字发表他自己的著作。因此，这一检验在戈塞特采用笔名"学生"发表后，正式被命名为学生 t 检验(student's t test)。

什么时候使用相关样本 *t* 检验(related *t* test)

所检验的关系类型	所要求的数据类型	研究设计
差异	等距	相关的: 配对组 重复测量

假设:数据来自正态分布总体。

优点:灵敏的检验——在使用威尔科克逊 *t* 检验得不出显著性时,使用它有时却能够得出显著性;使用了数据中更多的信息。

缺点:需要满足参数假设(parametric assumptions);计算复杂。

以后都将会有这样一个专栏来介绍每一种推断性检验。在一些 A 类的课程大纲中,你被要求回答选择某一特定检验的理由。你的策略应该是用以下信息中的两条或三条来回答,通常是:

- ❑ 所检验的关系类型(差异或相关)。
- ❑ 所要求的数据类型(称名数据、顺序数据、等距数据)。
- ❑ 研究设计:相关或非相关。

t 检验是我们所熟知的**参数检验**(parametric test)类别中的一种。这些检验建立在数据来自正态分布总体的假设的基础上。我们在第 10 章中已经讨论过这些内容。通常,检查数据以确保它没有明显的偏斜就足够了。在更高水平上,你需要通过计算偏斜度作更规范的检查。为使数据服从正态分布,需要在测量的**等距**水平上来搜集数据,这也是使用参数检验的一个必要条件。

我们测试的数据

假设我们呈现给 13 个被试一个有 20 个单词的单词表,让他们在两种条件下试着回忆这些单词。在第一种条件下,要求被试形成每一个单词的心理表象,而在第二种条件下,他们只是按指令在听到下一个单词之前复述每一个单词。在每一个单词表均被呈现后,被试的注意力被分散几分钟,然后他们被要求回忆单词表中的所有单词。这些条件都被平衡了,以抵消顺序效应。数据结果见表 11.1。

简要的解释

在这个被描述实验中,我们的期望是在表象条件下人们将做得更好。也就是说,我们期望与复述条件相比,他们在表象条件下能有所提高。我们如何测量提高的部分呢?我们只需要用他们在表象条件下所得的分数减去他们在复述条件下所得的分数。这个差异就是衡量他们提高部分的一个测量值。这些差异在表 11.1 第 3 列中被呈现出来了。如果在表象条件下丝毫没有优势(虚无假设),那么所有的这些差异都应该接近于零。一些略大于零,一些略小于零,而这些变异仅仅来自于抽样误差。试着设想一下,我们随机抽取一些差异并每次都计算他们的平均值。如果我们不断重复地做这样的工作,最后我们将得到如图 11.1 那样的一个分布。我们

希望我们得到的这些差异的平均数远远大于零，以给我们足够的把握去说明它的发生并不是偶然的。统计量 t 恰好可以告诉我们在虚无假设情况下，在多大程度上我们的平均数存在差异是不可能的。接下来，让我们来探讨一下 t 的计算。

表 11.1 表象和复述条件下正确回忆单词的数量

表象条件	复述条件	差异(d)	d^2
6	6	0	0
5	10	5	25
13	7	6	36
14	8	6	36
12	8	4	16
16	12	4	16
14	10	4	16
15	10	5	25
18	11	7	49
17	9	8	64
12	8	4	16
7	8	−1	1
15	8	7	49
		$\bar{d} = 4.54$	
		$S_d = 2.6$	
		$\sum d = 59$	$\sum d^2 = 349$

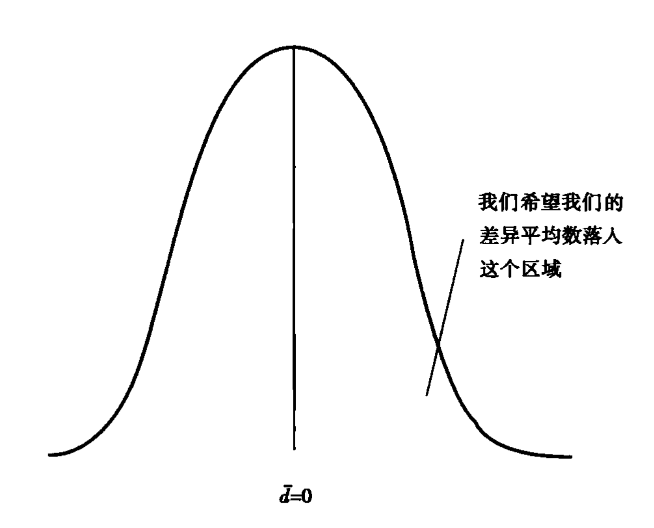


图 11.1 在虚无假设条件下多次计算 13 对差异的平均数

技巧提示：当你要在这样一个表格中（表 11.1）安排数据时，把你期待得分较高（如此例中的表象条件一列）的那一列放在其他得分组的左边是有意义的。那么，如果你的假设是正确的，将表中左边的数值减去右边的数值，你应该得到正值。记住，对于每对数据，你都必须始终在同一个方向上做减法。如果你在计算差异时，得

到了许多负值，那么你的结论与你所期待的假设恰恰相反，你得到的 t 值将是负值。这对显著性检验没有影响——但是只有当你使用双侧检验时，你才能得到显著性的差异。

相关样本 t 检验的计算公式

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{N}}}$$

这个公式要求我们找到差异的平均值，也就是 $\bar{d}$，然后除以标准差，而标准差本身已经除以了样本人数的平方根 $\sqrt{N}$。

相关样本 t 检验的计算步骤	用我们的数据进行计算
1. 找出差异平均数	$\bar{d}=4.54$　见表 11.1
2. 找出差异的标准差 = S（计算步骤详见表 8.7 或见下）	$s=2.6$　见表 11.1
3. 找出 $\sqrt{N}$	$N=13$ 所以 $\sqrt{N}=3.61$
4. 用 s 除以 $\sqrt{N}$	$\frac{s}{\sqrt{N}}=\frac{2.6}{3.61}=0.72$
5. $\bar{d}$ 除以第四步的结果得到 t	$t=\frac{\bar{d}}{0.72}=6.31$
6. 找出自由度（degrees of freedom）（df）*。对于一个相关样本设计，自由度等于 $N-1$，$df=13-1=12$	
7. 根据附录 2 中的表 3 来寻找 t 的临界值（critical value）。先找到你刚刚得到的 df 值所在的行，然后再查找在双侧检验下，“0.05”所在列的那个值。	你会发现相应的 t 值是 2.179。我们得到的 t 值**（obtained value）很容易就超过了 2.179 这个临界值。因此我们得到了一个可以拒绝虚无假设的显著结果。事实上，我们计算得到的 t 值已经超过了当 $p \leqslant 0.01$ 时的临界值 3.055。在这种情况下，一些研究者认为差异达到了极其显著的水平。

* **什么是自由度？** 这是一个回答起来有些复杂的问题。然而，它们的使用是相当简单容易的，当你要计算参数检验和卡方检验中的概率时，它们是必须的，这一点我们将在本章接下来的内容中提到。

** 也被称为“计算值”。

计算差异标准差

使用表 8.7 的步骤计算表 11.1 中 d 值的标准差要求可能有些高。作为替代，你可以使用下列公式并按下列程序计算：

$$S=\sqrt{\frac{\sum d^2-\frac{(\sum d)^2}{N}}{N-1}}$$ 在此要特别留意 $\sum d^2$ 和 $(\sum d)^2$ 之间的区别。

1. 如表 11.1 中那样计算 $\sum d^2=349$。

2. 计算 $\sum d$ 并求平均值 $= 59^2 = 3\ 481$。

3. 用第二步的计算结果除以 $N = \frac{3\ 481}{13} = 267.77$。

4. 用第一步的计算结果减去第三步的计算结果 $= 349 - 267.77 = 81.33$。

5. 用第四步计算结果除以 $(N-1) = \frac{81.33}{12} = 6.78$。

6. 计算第五步结果的平方根 $= \sqrt{6.78} = 2.60$。

当你必须使用并报告显著性检验时

使用并报告显著性检验对于刚接触心理学的学生来讲是非常棘手的操作，经常会引起极大的担心和迷惑。因此，在表 11.2 中，以我们刚刚进行过的 t 检验为例子，我提供了一个适用于进行**所有**显著性检验的“详细说明书”式的程序。如果你必须撰写一个研究报告，那么你在报告你的结果时需要特别谨慎。第 14 章将帮助你了解这些并引导你避免几个通常易犯的错误。通过这样或那样的途径，表 11.2 中使用的所有步骤，在你报告的结果部分都要提及。我强烈建议每次当你必须处理显著性检验时，最好参考表 11.2。报告刚才得出的 t 检验的结果的方法在专栏 11.1 中被呈现出来了。

表 11.2 选择、使用和报告一个显著性检验的步骤

选择一个适当的推断统计检验	第10 章将帮你了解这一点，通常在一个报告中你需要证明你的选择是正确的；当本书中的每一个检验都被介绍后，你可以根据给出的条件作出选择。
计算检验统计量	在我们刚才处理过的例子中，我们采用的是 t 检验，结果是 $t = 6.31$。所有的统计检验都要计算一个检验统计量，并用一个字母去表示它。
将查表所得的临界值与检验统计量作比较	这些表格在附录中。在这个例子中我们考虑以下几点：
思考：	
1. df 或 N——在每一个检验中，这个信息都将是被给出的。	1. 我们使用自由度(df)，对于相关样本检验来说，它等于 $N-1$，即 12。
2. 单侧或双侧检验	2. 我们使用双侧检验。
3. 确定适当的显著性水平	3. 通常是 $p \leqslant 0.05$，这也是我们上面所使用的。
决定你自己的结果在临界值的哪一侧。	当我们检查 t 时，我们发现它远远高于 0.05 和 0.01 显著性水平所需的临界水平。
表格将告诉你，你的检验统计量是否需要高于或低于临界值。	表格告诉我们，为了使差异被认为是显著的，t 需要高于临界值。
报告你的结论	既然 t 高于临界值，我们就有信心拒绝虚无假设(因为我们达到了 $p \leqslant 0.01$ 的水平)。专栏 11.1 显示了一个报告 t 检验结果的可以接受的途径。

专栏 11.1 t 检验分析结果的报告

> 正如预期的那样，在表象条件下回忆的单词量的平均数（$M = 13.38, SD = 3.52$）要高于在复述条件下回忆的单词量的平均数（$M = 8.85, SD = 1.68$）。平均数之间的差异是显著的，$t(12) = 6.31, p \leq 0.01$，双侧检验。
>
> 如果这个结果是不显著的，不要说它是可以忽略的结果。应该写作：
>
> ……差异是不显著的，$t(12) = 2.15, p > 0.05$。

独立样本 t 检验

什么时候使用独立样本 t 检验（unrelated t test）

所检验的关系类型	所要求的数据类型	研究设计
差异	等距	非相关的： 独立的样本/区组/处理

假设：数据来源于正态分布总体

方差齐性

优点：检验更灵敏——当曼-惠特尼 U 检验不能得出显著性的结论时，独立样本 t 检验可以；比非参数检验利用了更多的数据信息。

缺点：需要满足参数假设；计算复杂。

我们测试的数据

研究者要求被试连续写一个月的睡眠日记。他们区分出两组被试，一组记录了高水平的睡眠紊乱，另一组记录了低水平的睡眠紊乱。然后对这两组被试的焦虑水平作出测试，其相应的结果见表 11.3。

表 11.3 高水平睡眠紊乱组和低水平睡眠紊乱组被试的焦虑得分

高水平睡眠紊乱组被试		低水平睡眠紊乱组被试	
x_h	x_h^2	x_l	x_l^2
14	196	8	64
11	121	10	100
11	121	9	81
12	144	7	49
13	169	9	81
9	81	8	64
12	144	12	144
11	121	11	121
13	169	13	169
9	81		
Mean = 11.5		9.67	
$\sum x_h = 115$	$\sum x_h^2 = 1\ 347$	$\sum x_l = 87$	$\sum x_l^2 = 873$

下标 h 表示高水平（High），下标 l 表示低水平（Low）。

简要的解释

当我们要检验像第 9 章已经讨论过的两组螺丝钉样本时，这一检验是合适的。让我们回顾一下那一章，以便给出更充分的解释。现在我们所要再次强调的是，我们正在检验两组样本平均数之间的差异，而这两组样本被假设是来自同样的（或同一）总体。虚无假设是高水平睡眠紊乱组被试的焦虑平均分等于低水平睡眠紊乱组被试的焦虑平均分。我们通过假设差异存在并计算差异产生的概率的方法，来验证这一假设。现在我们做一下深呼吸，来看一看计算独立样本 t 检验 t 值的公式。

独立样本 t 检验 t 值计算公式

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\left[\frac{\left(\sum x_a^2 - \frac{(\sum x_a)^2}{N_a}\right) + \left(\sum x_b^2 - \frac{(\sum x_b)^2}{N_b}\right)}{(N_a + N_b - 2)}\right] \times \left[\frac{N_a + N_b}{N_a N_b}\right]}}$$

显而易见，这个公式看起来有些令人害怕！然而通过下面的方法，这个公式只是包含了一些简单的操作而已。然而如果你真的无法胜任这样的计算，我们还可以偷偷地作一个转换，即使用曼-惠特尼检验作为替代。

计算独立样本 t 值的步骤	使用我们的数据进行计算
1. 将高紊乱组所有得分相加；	$\sum x_h = 115$（参见表 11.3）
2. 将高紊乱组所有得分的平方相加；	$\sum x_h^2 = 1\ 347$（参见表 11.3）
3. 将第 1 步的结果进行平方[注意区分 $\sum x_h^2$ 和 $(\sum x_h)^2$]；	$(\sum x_h)^2 = 115^2 = 13\ 225$
4. 用第 3 步的结果除以 N_h—— 高紊乱组得分的个数。	$(\sum x_h)^2 / N_h = 13\ 225/10 = 1\ 322.5$
5. 用第 2 步所得结果减去第 4 步所得结果。	$\sum x_h^2 - (\sum x_h)^2 / N_h = 1\ 347 - 1\ 322.5 = 24.5$
6—10. 对低紊乱组的得分重复第 1—5 步的计算。	$\sum x_l = 87$（参见表 11.3）
	$\sum x_l^2 = 873$（参见表 11.3）
	$(\sum x_l)^2 = 7\ 569$
	$(\sum x_l)^2 / N_l = 7\ 569/9 = 841$
	$\sum x_l^2 - (\sum x_l)^2 / N_l = 873 - 841 = 32$
11. 将第 5 步和第 10 步的结果相加。	$24.5 + 32 = 56.5$
12. 用第 11 步的结果除以 $(N_h + N_l - 2)$。	$56.7 \div 17 = 3.324$
13. 用第 12 步的结果乘以 $(N_a + N_b)/N_a N_b$。	$3.324 \times (10 + 9)/(10 \times 9) = 0.702$
14. 求第 13 步结果的平方根。	$\sqrt{0.702} = 0.838$

续表

计算独立样本 t 值的步骤	使用我们的数据进行计算
15. 求两个平均数之差。	$\bar{x}_h - \bar{x}_l = 11.5 - 9.67 = 1.83$
16. 用第 15 步的结果除以第 14 步的结果得到 t 值。	$t = 1.83/0.838 = 2.183$!!!
17. 计算独立样本 t 检验的自由度(df),即 $N_a + N_b - 2$(注意我们在第 12 步中计算过了)。	$df = 10 + 9 - 2 = 17$
18. 参考附录 2 中的表 3,按 $df = 17$,双侧检验,$p \leq 0.05$ 查表。	所给出的临界值是 2.11,所以我们侥幸地通过检验,可以报告差异是显著的。

像此前给出的报告相关样本 t 检验的结果一样,报告该检验结果即可。

11.2 差异的非参数检验

接下来的两个检验,我们称之为非参数检验,因为它们不像参数检验那样要对潜在数据作出假设。我们不需要正态分布的数据,也不需要通过真正的等距量表来搜集数据。这种检验适用于当我们觉得我们所检验的数据不能被认为是真正的等距数据时。你也可以将这两个检验视作相关样本 t 检验和独立样本 t 检验的等价物,它们可以得到近似的结果,但更容易计算。

威尔科克逊配对符号秩次检验

何时使用威尔科克逊 t 检验

所检验的关系类型	所要求的数据类型	研究设计
差异	顺序变量(或更高水平)	相关的: 重复测量 配对

假设:仅适用于以上所列的条件。不要将这一检验与 t 检验混淆;它的统计量是 T(始终大写)。

优点:不需要满足参数假设。易于计算。

缺点:与相关样本 t 检验相比,对数据较不敏感,因此,当 t 检验能得出显著性结论时,它可能并不总是能得出显著性结论。

我们测试的数据

假设一所大学的某些讲师热衷于对某一单元进行改变,他们希望这一单元最终能通过基于计算机学习的方式传递给学生。这里没有传统的讲解和课堂讨论,所有的内容都经由电脑网络进行传递。为了检验计划的可行性,他们把这一单元的教学分成两部分。首先在第一学期用常规方法进行教学,接着在第二学期采用电脑教学。他们要求 15 个学生对每部分在态度量表上进行评分,分数为 10~50。具体数据见表 11.4。

表 11.4 学生对常规的和基于电脑的两种单元讲解方法进行评价的态度得分

常规呈现(A)	基于电脑的呈现(B)	两者差异(B—A)	差异等级	差异为正值的等级	差异为负值的等级
23	33	10	12	12	
14	22	8	9.5	9.5	
35	38	3	3	3	
26	30	4	5	5	
28	31	3	3	3	
19	17	-2	1		1
42	42	0	-		
30	25	-5	6		6
26	34	8	9.5	9.5	
31	24	-7	8		8
18	21	3	3	3	
25	46	21	14	14	
23	29	6	7	7	
31	40	9	11	11	
30	41	11	13	13	
			等级之和:	正值:90	负值:15

简要的解释

威尔科克逊检验适用于成对数据。也就是说,两组数据来自相同被试在两种不同条件下的得分或已相互匹配过的两组被试的得分。我们想知道的是,对于每一组数据,来自一种条件的得分是否有高于来自另一条件下的得分的倾向。在我们的例子中,如果基于电脑的课程被确定为对学生更有用,我们就会期望在每对数据中,基于电脑的课程的得分更高。如果我们考察每对数据的差异,总是用电脑的课程的得分减去常规课程的得分,大体上说,我们希望差异是正的,而负的差异是"不被期望"的——它们会削弱我们关于"基于电脑的方法更好"的假设。因为这是顺序数据的检验,我们将差异转换成等级——按照差异的大小排定差异等级,而忽略差异的符号(或方向)。然后我们将两组等级分别求和,一组差异为负数,另一组差异为正数。差异为负数不是我们所期待的,所以我们希望这一组的等级之和小一些。我们期待正数的等级之和大于负数的等级之和;我们想要尽可能少的负数等级出现。

使用的数据

我们将重新考察表 11.3 中的数据,这次我们将进行曼-惠特尼 U 检验。其原理与当一个团队里的每一个人分别与另一个团队里的每一个人进行比赛时,比赛将如何开展的原理一样。如果"鸭子和羽毛"标枪投掷队的全部 5 个成员击败了"白马"队的每一个成员,那么将不会有人敢说这全是因为运气所致。要一起进行 25 场比赛,因为一个队里的 5 个队员中的每一个队员都将与另一个队里的 5 个队员进行比赛。在 25 场比赛中,如果"鸭子和羽毛"队赢了 13 场,而"白马"队赢了 12 场,那么我们根本无法确认运气在这一结果中没有起作用。

数据在表11.5中再次被呈现出来，通览表格并对表中的每一个数据赋予点值。计算表中第2列和第4列的点值的规则如下：

- 每当一组中的一个分数被另一组中的一个分数击败时，就赋予这个数据1个点值。
- 每当一组中的一个分数与另一组中的一个分数打成平局时，则赋予这个数据0.5个点值。

威尔科克逊 T 值的计算步骤	使用我们的数据进行计算
1. 找出每对数据的差异，对所有样品，用我们期望它较大的数据减去我们期望它较小的数据。	参见表11.4
2. 忽略差异的符号，对所有差异排等级。忽视所有0。这些0将不被分析。	参见表11.4
3. 分别计算差异为正数和负数的差异等级之和。这两者中较小的那个就是 T 值。	差异为正数的差异等级之和 = 90 差异为负数的差异等级之和 = 15 $T = 15$
4. 用 N（这里不包括差异为零的样本）和适当的检验形式（单/双侧检验）代入附录2中的表4。**在这一检验中，你的值必须低于在表中找到的临界值。**	N 为14，使用双侧检验（他们可能更倾向于传统演讲的方法）；使用 $p \leq 0.05$，相应的临界值为21，我们的值更小。因此差异是显著的。我们可以断定我们的差异在0.02显著性水平上是显著的，因为我们的值等于在这一水平上的临界值。

专栏11.2 威尔科克逊检验分析结果的报告

> 一个学生对两种方法中的任何一种都没有显示出偏爱，被剔除出分析范围，这样 N 为14。赞同基于电脑的方法的等级之和为90，赞同常规方法的等级之和为15。威尔科克逊检验分析显示：$T = 15, p \leq 0.05$，采用双侧检验。

曼-惠特尼 U 检验

何时使用曼-惠特尼 U 检验

所检验的关系类型	所要求的数据类型	研究设计
差异	顺序变量	非相关的： 独立样本/群组/测量

假设：除了以上所列之外就没有有了。

优点：不需要满足参数假设。比 t 检验易于计算。

缺点：与独立样本 t 检验相比，对数据较不敏感，因此，当 t 检验能得出显著性结论时，它可能并不总是能得出显著性结论。

例如，高紊乱组中的第二个数据是11，它被另一组中的数据12和13击败（给予

2 个点值），同时还与另一组中的另一个数据 11 打成平局（给予 0.5 个点值）。因此其总点值为 2.5，这种计分系统的独特之处在于，你越差，你得到的点值越多；所以，很像高尔夫，其目标是使你所希望得到较好成绩的那一组获得较低的点值。这两个点值的总和中较小的那个被看做为曼-惠特尼 U 值。这里有 90 场竞赛（10×9），所以总点值必须是 90，实际情况也是这样。你可以通过计算竞赛的场数（$= N_1 \times N_2$）并始终确保 U_1 和 U_2 之和等于总竞赛场数。所以：

$$N_1 \times N_2 = U_1 + U_2$$

表 11.5 高紊乱组和低紊乱组被试的焦虑得分（曼-惠特尼检验）

高紊乱组被试		低紊乱组被试	
x	点值	x	点值
14	0	8	10
11	2.5	10	8
11	2.5	9	9
12	1.5	7	10
13	0.5	9	9
9	5	8	10
12	1.5	12	4
11	2.5	11	6.5
13	0.5	13	2
9	5		
	$\sum = 21.5$		$\sum = 68.5$

使用 U 值，每组的样本数，检验形式（单/双侧检验）和一个合适的 p 值水平，代入附录 2 中的表 5 检验 U 的显著性。如果我们说这个检验是双侧的，并且我们希望 $p \leqslant 0.05$，然后我们选择附录 2 表 5c。因为它对应于 $p \leqslant 0.05$ 的双侧检验。对于 $N_1 = 10, N_2 = 9$ 来说，这里给出的临界值是 20。需要注意的是，你把哪些被试称为 N_1，把哪些被试称为 N_2 都无关紧要。我们的 U 值是 21.5，所以我们差一点就可以认为差异是显著的。

专栏 11.3 关于曼-惠特尼检验分析结果的报告

> 每一组中的每个焦虑得分，当它们等于或低于另一组中的分数时，都被分配了相应的点值。较低的点值之和被当做是 $N_1 = 10$ 和 $N_2 = 9$ 时的曼-惠特尼 U 值。在低紊乱组中发现了较低的焦虑分数，但是差异不显著，$U = 21.5$，$p > 0.05$，双侧检验。低紊乱组的点值总数是 68.5，而高紊乱组的点值总数是 21.5。

参数检验和非参数检验的比较

这样的结果给我们带来了一些趣味，因为对同样的数据，使用独立样本 t 检验，我们得到了差异显著的结论。在这种情况下，与非参数 U 检验相比，参数 t 检验显

示了其更灵敏的特性。在96%的情况下，t 检验与 U 检验或 T 检验将得到相同的结果。参数检验具有更大的效力，因为它们使用了更多可以利用的数据。

练 习

1. 找出表格中 a，b，c 和 d 四个例子中的检验统计量是否具有显著性。并给出在单/双侧检验两种情况下从表格中所能得到的 p 的最小值。每组的样本数和合适的检验方法已经给出。你可以将 p（最小概率值）写在"显著性"这一列的空白处。

	每组的样本数		曼-惠特尼检验	显著性			威尔科克逊检验		显著性	
	N_a	N_b	U	单侧	双侧		N	T	单侧	双侧
a	15	14	49			c	18	35		
b	8	12	5			d	30	48		

2. 使用下表中的 t 值，并检查其显著性，完成表格中的最后两列。

	t	N	研究设计	单/双侧检验	$P\leq$	拒绝虚无假设?
a	1.75	16	相关	1		
b	2.88	20	非相关	2		
c	1.7	26	非相关	1		
d	5.1	10	非相关	1		
e	2.09	16	相关	2		
f	3.7	30	相关	2		

3. 选用双侧检验，使用表11.1的数据进行适当的检验（曼-惠特尼检验或威尔科克逊检验），检查其结果的显著性。
4. 浏览下列表 a 和表 b 中的数据，指出每个例子是否适合进行 t 检验，并对该检验的假设作出检查。

a.			b.		
	17	23		17	23
	18	9		18	11
	18	31		18	24（相关数据）
	16			16	29
	18	（非相关数据）		12	19
	17			15	16
	6				

同时：为每组数据选出合适的非参数检验。计算出每组数据的 t 值和相应的非参数检验的统计量。

5. 一个研究报告显示，在一个样本容量为11的重复测量设计中，t 值为2.85，$p<0.01$，即结果具有显著性。这一假设检验是否有可能采用了双侧检验？

答 案

1. a. 0.01（单侧），0.02（双侧）；b. 0.005（单侧），0.01（双侧）；c. 0.025（单侧），0.05（双侧）；d. 0.001（单侧）；0.002（双侧）。
2. a. NS，保留 H_0；b. 0.01，拒绝 H_0；c. NS，保留 H_0；d. 0.005，拒绝 H_0；e. NS，保留 H_0；f. 0.01，拒绝 H_0。
3. 威尔科克逊 $T=1$；$p<0.002$。

4. a. 方差根本不齐性，非相关设计或样本数差别极大。因此进行 t 检验是不明智的。相应的非参数检验：曼-惠特尼检验。b. 虽方差不齐性，但是是相关设计。因此采用 t 检验是安全的。相应的非参数检验：威尔科克逊符号秩次检验。a. t(非相关设计)1.14；b. t(相关设计)1.57。a. $U=6$；b. $t=4.5$。
5. 不可能。自由度(df) = 10。$P\leq0.01$ 时的双侧临界值 = 3.169。

关键术语

临界值(critical value)	显著性检验(significance test)
推断检验(inferential test)	t 检验；相关的；非相关的(t test；related；unrelated)
曼-惠特尼检验(Mann-whitney test)	检验统计量(test statistic)
实际值(obtined value)	威尔科克逊配对检验(Wilcoxon Matched pairs test)

11.3 分类数据的检验

卡方检验(写作 χ^2，读作"卡方")

何时使用卡方检验

所检验的关系类型	所要求的数据类型	研究设计
关联或差异	称名变量	非相关的

假设：(详见本节末尾)

样本(通常是人)只可能出现在交叉数据表格的一个单元格中。

较低的期望频数是个问题。

优点：对于独立分类这一水平的数据，这是唯一可以被很好地接受的检验。

缺点：参见假设以及下面关于低期望频数和可供使用的数据类型的详细论述。

使用的数据

当数据以频数的方式搜集时，可以使用卡方检验。通常它们像表 11.6 中那样排列。一般来说，你依据某一变量将人(或者是像例子中所提及的汽车)进行分类，你需要明白的是，这些分类与其他类别变量的分类是否有本质的不同，也就是说，数据在分类中变成了频次。因为每一个被观察的对象可以是物(比如汽车)也可以是人，所以，我们通常把被观察的对象称为样品。在此水平上，在实际工作中观察不同性别间的差异是很普遍的。典型的实践考察包括观察男性和女性在以下行为中是否有差异：他们拿书的方法(夹在腋下还是放在胸前)，停放车辆的方式(向前进还是向后倒退)，当两个人在狭窄的地方碰面时各自采取的通过的方法(面对面还是背对背)，自我推销时的方法，等等。每一次的分类变量，第一个都是性别，第二个则是观察到的行为。为了使用卡方检验，被测行为必须是**分类**的，而不是记分的。接下来，我们以具体数据为例——比如按吸引力为某人打分，从 1 到 10——然后通过将这些数据分类，把它们降格为称名变量，比如，分为"高""中"和"低"。

表 11.6 旧车和新车在黄色交通灯前停车或不停车的频数

在黄色交通灯前的行为	车辆的类别		
	新车	旧车	总计
停车	90_a	88_b	178
不停车	56_c	89_d	145
总计	146	177	323

表格 11.6 中的频次数据的排列就是我们熟知的交叉表(cross-tabs)。在做卡方检验时,除了你自己出于简洁和不易混淆的考虑而设置的限制之外,表格中的行和列的数目不受限制。例如,如果我们在交通灯前观察到更多关于汽车和驾驶员行为的细节的话,表 11.6 中的数据就可以像表 11.7 中那样排列。

表 11.7 关于表 11.6 中数据的更详细的描述

在黄色交通灯前的行为	新车(使用不足 2 年)	半新的汽车(2~5 年)	旧车(使用 5 年以上)
停车	65_a	68_b	45_c
减速	35_d	18_e	20_f
保持相同的速度	15_g	26_h	31_i

在某些情况下,只要有一列数据就可以做卡方检验,就像表 8.1 一样,如果虚无假设正确,只要将这些数据的实际排列情况和他们本来应该如何排列相比较。同时也可以参见表 11.8,以便获得更充分的解释。

一些说明

假设你从某处了解到性格外向的人就是那些不会为社会习俗所困扰的人,他们不介意大胆地展示自己。如果这是正确的,那么性格外向的人不会介意在沙滩上赤身裸体,事实上,他们非常喜欢裸泳。做这样一个试验是多么聪明的主意啊! 你兴冲冲地跑到一个众所周知的对裸体主义者很宽容的希腊(Greek)小岛上,这里阳光普照。然后对沙滩上的每一个人进行外向性和内向性的测试。我并不推荐这种测试方法——在行动之前与你的导师检验一下伦理标准(见第 13 章)!

假设你决定,当某人的分数高于平均数时就被定义为"性格外向的人",而"性格内向的人"则是指分数低于平均数的人。你所获得的这些数据可能出现在表11.8中观察频数(observed frequencies)这一行中。在此我们能看到有 40 个赤裸的性格外向的人和 10 个赤裸的性格内向的人。在这里,我们所检验的虚无假设就是性格内向的人和性格外向的人在裸露方面没有差异。如果这一假设成立的话,那么可能脱掉衣服裸露的性格内向的人的数目和性格外向的人的数目保持一致。在这种情况下,假设这个虚无假设是正确的,那我们就期待在所有样本中一半是性格内向的人,一半是性格外向的人。也就是说,在 50 个赤裸的人中,应该有 25 个是性格外向的人,25 个是性格内向的人。这是我们的预期,因此我们把它称为"期望频数"(expected frequencies),参见表 11.9。如果虚无假设是正确的,期望频数就是我们在样本中所期望得到的东西。卡方检验考查的是"我们所观察到的"与"在虚无假设下我们所期望的"这两者之间差异的大小。

表 11.8 性格外向的人和性格内向的人赤裸的频数

	性格外向的人	性格内向的人	总计
观察频数	40_a	10_b	50

表 11.9 与表 11.8 相关的性格外向的人和性格内向的人赤裸的期望频数

	性格外向的人	性格内向的人	总计
期望频数	25_a	25_b	50

计算卡方

要进行卡方检验分析，首先要把观察频数中的每个单元格用一个字母来标识。单元格就是指原始数据表格中的一个方框（但不包括“总和”这一格）。在表 11.8 中，只有两个单元格，因此把这两个单元格标为 a 和 b。期望频数的相应单元格需要用相同的字母。

卡方的计算公式

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O 是观察频数一行中每个单元格的值，而 E 是每个相应的期望频数。这个公式要求做一组关于每一对观察频数和期望频数的计算，然后再像下面这样把结果相加。

	$O-E$	$(O-E)^2$	$(O-E)^2/E$	结果
性格外向的人	40 − 25 = 15	225	225/25	9
性格内向的人	10 − 25 = −15	225	225/25	9
			$\chi^2 =$	18

对于每个观察表格，我们先用观察频数（O）减去期望频数（E），再将得到的结果进行平方。然后用这个结果除以期望频数。用这种方法依次对每个单元格进行计算，最后将得到的结果相加，就像上面计算表格中右边一列一样。

检查表中的结果：在我们检查 χ^2 结果的显著性之前，首先需要找出自由度。在这一特殊的例子中，χ^2 的自由度是观察频数单元格的数目减去 1，所以我们得到 2 − 1 = 1。在这种由两个数据单元格组成的表格的情况下，我们可以用单侧检验。除此之外的 χ^2 检验的其他所有情况都必须使用双侧检验。参考附录 2 中的表格 6，我们可以得知，当 $p \leq 0.05$ 时，χ^2 值需要大于 2.71。事实上，当 $df = 1$ 时，我们的 χ^2 值都大于表格中的所有其他水平下的值，所以这个结果是极其显著的。因此，对于性格外向的人更有可能在沙滩上赤身裸体的假设，我们给出了强有力的支持。

或者我们可以……

我们上面所做的只是一个关于“性格外向的人比性格内向的人更有可能在沙滩上脱光衣服”的简单的假设检验。我们仅仅使用了那些在沙滩上裸露的人做样本。我们忽视了其他人。假如内向的人只是不愿意去沙滩，情况又会怎样呢？假设现在有 50 个穿着衣服的人在沙滩上，其中我们发现 10 个人是内向的，40 个人是外

向的。那么我们知道裸露的内向人的比例与沙滩上的内向人的总比例是相等的，这样我们就不能很好地支持以下假设：内向的人没有外向的人倾向于裸露。我们要做的是找出在沙滩上裸露的和穿衣服的外向人和内向人各自的比例，以便作出公平合理的比较。现在我们有两个变量，都是类别变量。一个是含有两个水平的人格变量，外向和内向。另一个是穿着状态变量，裸体的和穿衣的。如果我们发现我们整理的数据如表 11.10 所示，那么我们就可以作出一个更为公正的结论，内向的人更喜欢保持穿着衣服的状态。

表 11.10　沙滩上裸体的和穿衣的外向人和内向人各自的观察频数

人格类型	穿着状态		总计
	裸体的	穿衣的	
外向的人	40_a	10_b	50
内向的人	10_c	40_d	50
总计	50	50	100

我们发现（为了解释起来更方便）尽管沙滩上的内向人和外向人一样多，但与 50 个外向人中有 40 个人裸露相比，50 个内向人中只有 10 个人裸露。

计算交叉表中的期望频数

如前所述，卡方检验分析的是观察频数和虚无假设情况下的期望频数之间的差异。在这里，虚无假设是内向人和外向人裸体频率没有区别（至少对于沙滩上的人是这样）。如果是这样，那么我们将期望表 11.10 的单元格中出现什么样的情况呢？我希望你同意以下说法：如果一半人是内向的，一半人是外向的，并且如果内向人和外向人之间没有区别，那么我们期望一半的内向人是裸体的。实际上，我们将期望得到如表 11.11 中所示的频数。

表 11.11　沙滩上裸体的和穿衣的外向人和内向人各自的期望频数

人格类型	穿着状态	
	裸体的	穿衣的
外向的人	25_a	25_b
内向的人	25_c	25_d

如果你同意 50 个内向人中应该有 25 个人是裸体的，那么你可以在头脑中以某种方式使用以下计算公式：

期望频数（对于每一个单元格而言）$=\frac{RC}{T}$，这里的 R 是单元格所在行的数值之和，C 是单元格所在列的数值之和，T 是所有单元格内数值的总和。在这个例子中，对于每一个单元格 a,b,c 和 d，我们都可以得到 $\frac{50\times 50}{100}=25$。

计算 2×2 交叉表的 χ^2 值

我们将利用表 11.6 中的数据计算其 χ^2 值，这项任务并不简单，这与你即将要处理的数据类别很相似。

步　骤	使用我们的数据进行计算
1. 对每个观察数据单元格用一个字母来表示。	参见表 11.6
2. 利用先前提供的公式计算期望频数。	单元格 a: $E=\frac{146\times178}{323}=80.46$
	单元格 b: $E=\frac{177\times178}{323}=97.54$
	单元格 c: $E=\frac{146\times145}{323}=65.54$
	单元格 d: $E=\frac{177\times145}{323}=79.46$

3. 利用前面提到的卡方的计算公式计算 χ^2 值，如下：

	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$	结果
单元格 a	$90-80.46=9.54$	$9.54^2=91.01$	$\frac{91.01}{80.46}=$	1.13
单元格 b	$88-97.54=-9.54$	$(-9.54)^2=91.01$	$\frac{91.01}{97.54}=$	0.93
单元格 c	$56-65.54=-9.54$	$(-9.54)^2=91.01$	$\frac{91.01}{65.54}=$	1.39
单元格 d	$89-79.46=9.54$	$9.54^2=91.01$	$\frac{91.01}{79.46}=$	1.15

$\chi^2=$（结果的总和）$=4.6$

4. 利用公式 $(R-1)(C-1)$ 计算自由度(df)，这里的 R 表示行数，C 表示列数。$df=(2-1)\times(2-1)=1$。
5. 使用双侧值和适当的 df，参考附录 2 中的表 6，检验 χ^2 的显著性。注意除了以前提到的只有一行两个单元格的特殊情况外，所有 χ^2 检验都必须使用双侧检验。
   我们有 1 个自由度(df)。表中给出的临界值是 3.84，我们计算出的 χ^2 值必须超过这一数值。4.6 远大于 3.84，所以我们在 $p\leq0.05$ 水平上得到差异显著的结论。

专栏 11.4　卡方检验结果的报告

> 177 个旧车司机中有 89 个在黄色交通灯前不停车，而 146 个新车司机中只有 56 个人没有停车。关于新车和旧车在黄色交通灯前停车与不停车的频数的 χ^2 检验是显著的，$\chi^2(1,N=323)=4.6, p<0.05$。看来与新车相比，更多的旧车在琥珀色交通灯前不停车。

简捷的 2×2 交叉表计算公式

这一公式只有在 2 行 2 列的情况下才能使用，正如上例一样。它节省了计算期望频数的劳力，如果你方便使用计算器，你会发现这些都可以使用单元格里的数据一步完成。

$$\chi^2=\frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

这里 N 代表总的样本容量，a，b，c 和 d 代表相应单元格内的数值。

更加复杂的交叉表

上面提及的详细的计算步骤(不是简捷公式)可以推广到更为复杂的表格,就像表11.7。这里的自由度是$(R-1)(C-1)=(3-1)\times(3-1)=4$。如果你愿意的话,你可以像做习题一样计算出其$\chi^2$值,也就是15.99。这个值超过了临界值9.49($p\leq0.05$, $df=4$,双侧检验)。

拟合优度(goodness of fit)的χ^2检验——来自单一变量的一行数据

当你掌握的数据仅仅是一个变量上的几个类别时,可以采用χ^2检验——就像我们处理的海边裸体的性格外向的人的人数一样。我们只有一个变量——人格类型:内向或外向——因此对于这个变量我们只有2个单元格。假设现在要搜集在四堂不同的数学课中冬季缺课学生人数的数据,这四堂课是由四位不同的导师教授的。每个导师教30位学生,每个学生只能在一个教室中上课。我们来看一下表11.12中的数据,这里,我们可能会怀疑雷基(Ricky)所带的班级存在问题。如果你确实得出这样的数据,那么你可以作出一个潜在的虚无假设。这里我们可以假定任何一个教师的课都没问题,因此他们课堂上学生缺席的频数应该是相同的。在某些情况下,你可以就期望频数提出不同的建议——比如,如果班级里的人数不相等的话,所有的缺席人数应当和每个班级里的学生数成比例,而不像我们这里举的例子。

表11.12 睿稚大学(Everlearn College)冬季学期不同数学教师所教课程中学生缺课人数

	数学导师的姓名				总计
	菲尔	舒利	雷基	迈克	
冬季学生缺课人数	5	2	9	4	20

这个表格的χ^2值的计算和我们在本节开头提到的以双向表来计算裸体的内向人和外向人的方法是完全一样的,只不过在计算中我们必须要计算单元格a,b,c和d(并不仅仅是a和b)。按照前面的虚无假设,缺席人数的期望频数应该随机地分布在各个单元格中。我们有20个缺席的人和4个单元格,因此,用20除以4得到5,这就是每个单元格的期望频数。那么自由度就是比类型数少1(对所有的单行检验都是如此),也就是3。你可能更愿意计算χ^2值并从表格中解释其结果。如果是这样,在这里你应该发现:$\chi^2(3, N=20)=5.2$, *NS*。"*NS*"表示"不显著"。$p\leq0.05$,自由度为3时的临界值为7.82,而我们得到的χ^2值没有达到这个水平。

使用卡方检验的限制

1. **一个个体必须仅仅出现在一个单元格中**。如果有学生在一个以上的班级中注册,那么在表11.12中他们可能缺席的频数可能同时出现在两个单元格中。这将使分析变得无效。当你询问人们他们参加什么活动并计算每项活动的参加人数时,类似的情况就会发生。如果你同时将某人记录为游泳和骑自行车,他们将出现在一个以上的单元格中,那么这样的数据是不能做卡方检验的。
2. **单元格中只能出现频数**。如果单元格里包含百分数、平均数、比率或除了单一事件的计数数据之外的其他任何数据,那么卡方检验是不能进行的。

3. **低的期望频数是一个问题**。如果超过20%的单元格的期望频数小于5，你的卡方检验结果就是不可靠的。因此在2×2表格中，既然每一个单元格占到所有单元格数的25%，那么任何一个单元格的期望频数小于5都将是一个问题。有些统计学家认为这一规则过于苛刻，关于这一问题的更多细节库里坎作了更详细的说明(Coolican,2004)。人们普遍认为要是样本大于20，且期望频数中只有一个或两个小于5时，此时继续做下去尚是比较合适的。然而，最好是设计好你的研究，搜集足够多的数据，以免陷入这样的困境中。

在一个单元格中出现的频数非常少的典型设计，往往出现在对一些相对稀少的人或人群的调查中。比如说左撇子、有阅读障碍的人、信仰特殊宗教的人（在这个国家中）、喜欢打人的人或者4岁以下知道液体守恒的儿童。要避免在实验设计中，有一个类型很难找到对应的人，应该利用大量的时间和资源从大量的人群中获取大量的数据。

二项符号检验

什么时候使用符号检验(sign test)

所检验的关系类型	所要求的数据类型	研究设计
差异	称名变量	相关的

假设：没有，且保证数据是成对的。

注意：对于每对数据，只有你确实可以说被试是"提高了"（或朝一个方向发展），"更差了"（朝另一个方向发展）或是"保持不变"时，才能使用符号检验。

优点：对于这一水平的相关类别数据，这是唯一被认同接受的检验。

缺点：对加工过的数据利用不够，不灵敏。在其他更灵敏的检验能得出显著性结论时，它经常不能得出显著性结论。

有时，会有这样一些数据，这些数据对于每一个人来讲，我们可以说它朝一个方向发展，朝另一个方向发展或不发生改变。如表11.1中的数据我们就可以这样说。我们可以仅仅通过标明每一个被试在第二种条件下是进步了（标作"+"），更差了（标作"-"）或者是保持不变（这样的个体不参加分析）来进行一项符号检验。然而，这并不是最明智的选择。虽然符号检验操作起来很容易，但用这种方法处理结果会使我们丢失太多原始数据的信息，我们经常不能发现本来就存在的显著性。我们会犯更多的第Ⅱ类错误。

我们测试的数据

然而，在儿童被要求参加一项皮亚杰守恒任务这一情景中，我们所记录的唯一信息是他们守恒或不守恒。首先，我们用常规方法，通过询问儿童两杯橘子汁是否有相同的数量来对他们进行测试。当他们认为相同时，他们是守恒的，然后我们将一个杯子里的橘子汁倒入另一个形状不同的杯子里。然后询问他们是否一个杯子有更多、更少或等量的橘子汁。如果儿童守恒，他们说"等量"。在第二种测试中，我们省略了询问他们两个相同的杯子里是否有等量的果汁这一步。这是为了避免给儿童以这样的暗示：如果被问及两个问题，"肯定有些不同"。当然，做这两个试

验在实践上是存在问题的，不过，让我们假设这已经在设计中作了解释，并且如果第二种方法（只提“一个问题”的方法）是更为公平的询问方法并更容易显示儿童能对液体体积守恒，那么看看我们希望从我们的儿童那里得到什么样的结果。

这一假设的研究数据见表 11.13，在这里，大部分儿童在传统测试中不守恒，而在“一个问题”的试验中却守恒，用一个“+”表示。当一个儿童一开始守恒，然后不守恒（这个儿童变得“更差”了），我们用一个“-”表示。当儿童在两个试验中做出相同的反应时，我们用“0”表示，并把他从最后的分析中剔除出去。

表 11.13 儿童在两种不同液体体积测试中的守恒性成绩

儿童	常规方法	一个问题的方法	符号
A	N	C	+
B	C	N	-
C	N	C	+
D	C	C	0
E	N	C	+
F	N	C	+
G	C	N	-
H	N	C	+
I	N	C	+
J	N	C	+

C = 守恒；N = 不守恒；+ = 仅仅在第二次守恒；- = 第一次守恒然后不守恒；0 = 在两次试验中出现相同成绩。

简要的解释

符号检验简单地假设，如果虚无假设是正确的，那么“+”和“-”号的数量应该是相等的。这正是我们在处理第 8 章中提到的关于“婴儿性别推测”的数据时所使用的原理和程序。我们希望不出现负号。如果虚无假设正确的话（即负号和正号的数量可能是相等的），让我们来计算一下我们实际得到的负号的数量的概率。如果这一概率低于 0.05，我们就有证据支持“一个问题”的试验方法提高了儿童守恒性的说法（或者不如说，“两个问题”的试验方法抑制了儿童的守恒性）。

二项符号检验中 S 的计算

1. 对于每对数据，依次赋予朝一个方向发展（通常是提高）的一对数据一个“+”号，赋予朝相反方向发展（通常是更差）的一对数据一个“-”号，或赋予没有变化的那对数据“0”。
2. 将标记为“0”的样品从分析中剔除出去。
3. 计算出现次数最少的符号的数目，我们将它定为 S 的值。在上面的例子中，负号出现的频率更小，即 2 次，所以，$S=2$。
4. 使用附录 2 中的表 7 查出临界值。我们得到的 S 值必须小于这个临界值。对于我们这个例子来说，采用双侧检验，$p \leqslant 0.05$，$N=9$（注意我们将儿童 D 剔除了），临界值是 1。很不幸，我们的 S 值是 2，不小于 1，所以结果是不显著的，在这一研究中我们不能认为“一个问题”的试验方法与“两个问题”的试验方法有所不同。我们差一点得到了差异显著的结论，所以用更多的被试重复这一研究是明智的。

专栏11.5 符号检验结果的报告

对于每一个儿童，当他们第一次不守恒而第二次守恒，我们将他记为正号；一开始守恒然后不守恒，我们将他记为负号。一个儿童在两个试验中都是守恒，这个结果被剔除出去了。使用保留下来的9个结果进行符号检验，发现正号和负号之间差异不显著，$S=2$，$p>0.05$，NS。

练 习

1. 在一项调查中，《卫报》的读者中有17人赞成狩猎狐狸，48人持反对意见，而在《每日快报》的读者中有33人赞成，16人反对。用这些数据进行卡方分析并对其结果进行解释。
2. 下面的数据可以使用卡方检验吗？

	守 恒	不守恒
4岁儿童	1	7
6岁儿童	7	1

3. 在另一项调查中，40个人对禁止售烟表示支持，而60人表示反对。
   (a) 为了检验支持售烟的人数是否具有显著多数性，用哪种类型的卡方检验来分析这些数据？
   (b) 计算 χ^2 值并对结果进行解释。
   (c) 这个检验可以使用单侧检验吗？
   (d) 如果另一调查显示74%的人反对禁止售烟，而26%的人赞成，可以用卡方检验这些数据吗？
4. 9个人参加了人际关系技巧的训练课程。在参加前和参加后他们都在其公司被问及参加这种类型的课程是否有必要。与参加课程之前相比，在参加完课程后，7人认为其必要性不如他们参加课程前所预期的，1人认为更有必要，还有1人的看法没有改变。
   (a) 哪个检验可以告诉我们，这门课程表面的负面影响是否有显著性？
   (b) 计算检验统计量并解释其结果。

答 案

1. $\chi^2(1, N=14)=19.25, p<0.001$。
2. 不能。所有的期望频数都小于5且不足20个被试。
3. (a) 一行两个单元格的卡方检验。
   (b) $\chi^2(1, N=100)=4, p<0.05$（刚刚好），拒绝虚无假设。
   (c) 可以——但仅仅只能在这一种情况下使用。
   (d) 不能，因为这些是百分数，只有频数才能使用卡方检验。
4. (a) 符号检验。
   (b) $S=1, N=8$（因为没有变化的1人被排除），所以临界值是0，但是我们得到的值是1，所以，参加课程前后对该课程的评价之间的差异并不显著。

关键术语

卡方检验(chi-square test)	拟合优度检验(goodness of fit test)
交叉表(cross-tabs table)	观察频数(observed frequencies)
期望频数(expected frequencies)	(二项)符号检验((binomial) sign test)

12

相关
——事物共同变化的趋势

本章内容

本章集中描述了与相关有关的概念和规律。

- 相关度衡量的是密切联系的两个变量共同变化的程度。如果一个变量随着另一个变量的增加而增加,我们称之为正相关;反之,则属负相关。
- 强度是衡量相关程度的指标,它的取值介于 -1 到 +1 之间,其中, -1 代表完全负相关; +1 代表完全正相关;0 代表没有相关。
- 显著性检验评估零假设下出现相关的可能性。(零假设总体相关为0。)
- 如何通过散点图来描述相关。
- 两种主要的计算相关的方法:
  ——**皮尔逊积差相关**:适用于等距数据。
  ——**斯皮尔曼等级相关**:皮尔逊相关在等级数据集上的应用。适用于顺序数据。
- 相关的几个关键点:
  ——即使两个变量之间存在强相关,我们也不能由此推断出它们之间存在因果关系。
  ——相关的两个变量不能是分类变量。

在日常生活中你可能对相关的概念已经非常熟悉。在功课上你越努力,可能得到的分数就越高。在孩童时期,你知道年龄越大,你就会长得越高。相关是两个变量之间的关系。假设你开了一家冷饮店。气温越高,你的销售额就越高。在销售额和温度这两个变量之间有一个关系。当想到相关的时候,我们通常趋向于只关注关系的一个方向,即天气越热,销售额就越高。但在你的冷饮店里同样存在这样一个事实,即温度越低,销售额就越低。这个相关是存在于温度和销售额这两个可被操作的变量之间的。这里有更多的可能相关的变量,这些变量以可测量的术语来描述。你可以通过在某个研究里测量这两个变量来检验可能存在的相关关系。

表 12.1 被推荐的相关和恰当的变量

可能的关系	变 量
1. 你给孩子们的越多,他们的期望就越高。	资源,期望值(例如猜想可以得到的圣诞礼物)
2. 个子越高,他们在事业上就越成功。	身高,成功(例如到达的水平,工资)
3. 我们的年龄越大,记忆就越差。	年龄,回忆的成绩
4. 天气越热,人们的攻击性越强(见图 12.1)。	温度,攻击性(例如攻击,脏话)
5. 你练习吉他的次数越多,所犯的错误就越少。	练习时间,错误
6. 智力来源于家庭。	父母的智商,孩子的智商

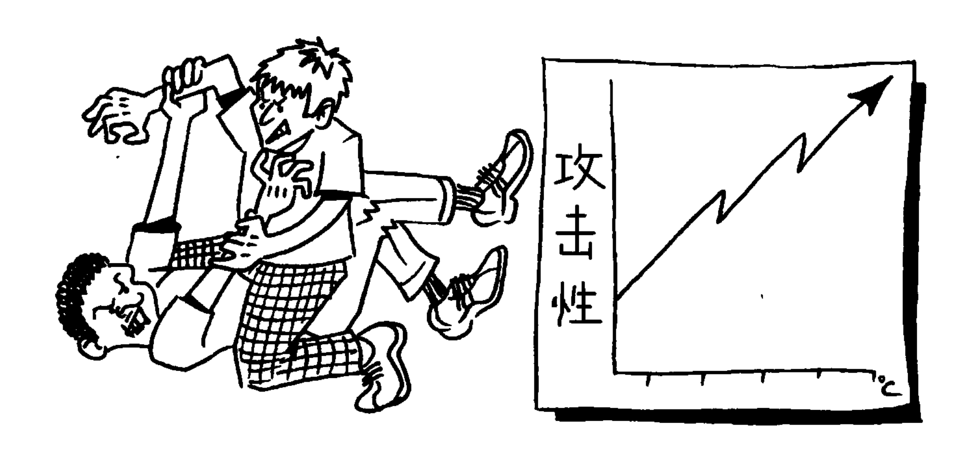


图 12.1 温度越高,人们的攻击性越强

12.1 正相关和负相关

在一个**正相关**(positive correlation)里,如果一个人在一个变量上得分很高,那么他在另一个变量上的得分也会很高。同样地,如果他在一个变量上得分很低,则在与它相关的另一个变量上得分也很低。举个例子:

你要送的报纸越多,所花费的时间就越多。

从心理学的角度,我们可以预期自尊(self-esteem)与自信之间存在正相关。在自尊上得分高的人往往在信心上的得分也很高。在自尊上得分低的人,在信心上得

分也很低。

在一个变量上的值是高的，但是在另一个变量上的值却是低的，我们称这种相关为**负相关**(negative correlation)。例如：

你要携带的纸越多，你走的就越慢。

在心理学中有一个例子：自尊与焦虑之间可能呈负相关。那些在自尊上得分高的人可能在焦虑上得分比较低；**反之亦然**。在图 12.2 里，距离 d_1 和 d_2 之间的相关是完全负相关；当一个高时，另一个就低；**反之亦然**。

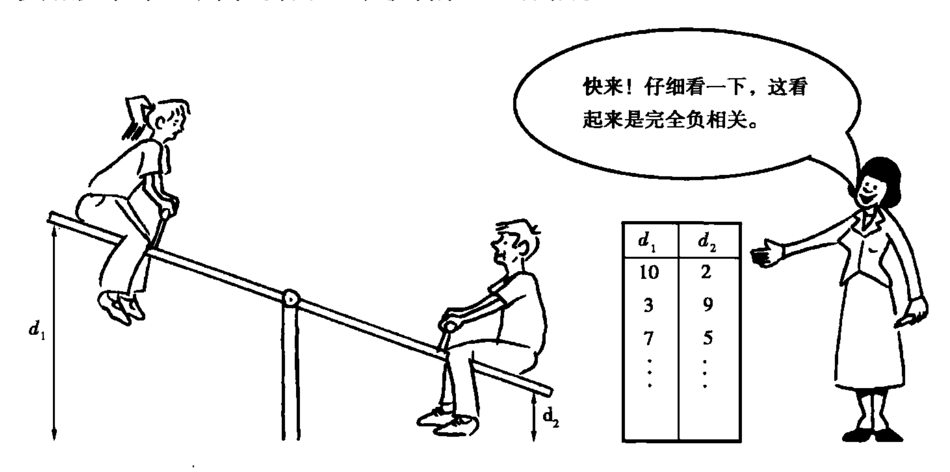


图 12.2 此消彼长的完全负相关

12.2 测量相关的强度

为了测量相关，我们必须有一对测量工具，即在相关的两个变量中，每一个变量都必须有一个测量工具。表 12.2 里的数据是典型的：对每一个被试来说，在自尊和信心上都各有一个分数。我们将看到两种计算相关的方法，但是每一种方法，我们都要用到一个被称作**相关系数**(correlation coefficient)的值。正如图 12.3 所表示的那样，相关系数总是一个介于 -1 和 +1 之间的值。注意，在一场考试里你可能会被要求描述所给的那个值的相关水平——例如"强"(strong)，"中等程度"(moderate)或者"弱"(weak)。

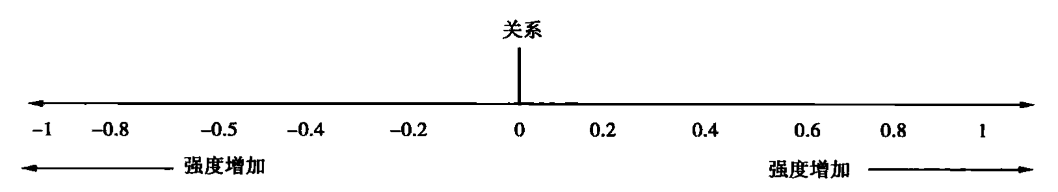


图 12.3 从完全负相关到完全正相关的测量

把一个负相关的值看做强相关可能是奇怪的，例如 -0.9。在这里，它是指按照上面的解释，相关是**负向的**，但是，随着相关系数从 -0.1 到 -0.9，一个变量的分数

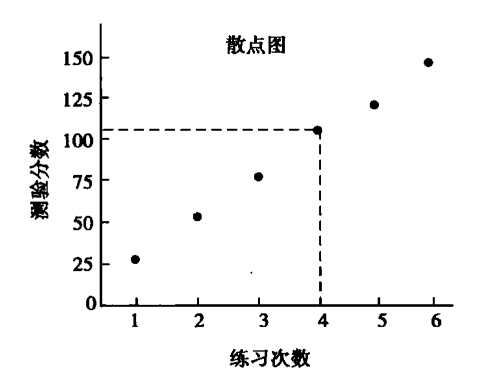
增加与另一个变量的分数减少的负向相关程度增加了。

得到一个比 +1 大或者比 -1 小的相关系数是不可能的。如果你发现自己得出的相关系数是这样的结果，那么你需要检查一下你的计算方法。因为你可能已经在某个地方或者其他地方犯了一个错误（或者你可能给你的计算机或者计算器下达了错误的指令）。一个值为零的相关系数告诉我们这两个变量之间完全没有关系。例如，我们可以预期一个人出生的时辰与他在 16 岁生日宴会上吃的鸡肉的数量之间没有关系。当我们检验一个相关系数的显著性时，我们的零假设为总体相关为零。

考试小贴士：我们很容易认为"负相关"代表着"没有相关"。你需要记住：没有相关意味着完全没有关系，但是负相关意味着存在一个相关，只是一个变量上得分高，与它相关的另一个变量得分低，**反之亦然**。

12.3 散点图——相关的图像

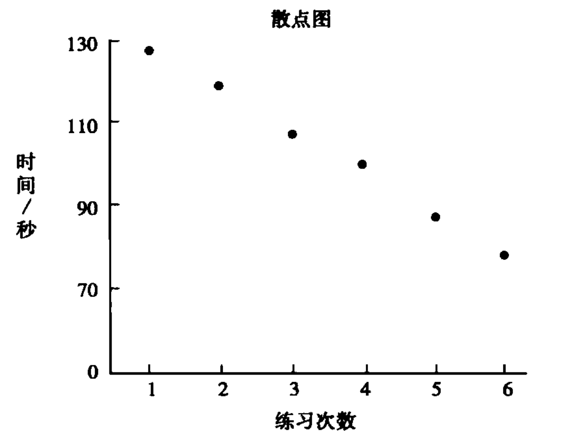
把相关形象化的一种有效方法是画一个**散点图**（scattergraph）（也叫做"scatter gram"，"scatter plot"）——也许在学校的某段时间里你可能已经做过散点图，用来表示鞋子的尺寸与身高之间的关系，那是老师最喜欢看到的事情。看图 12.4，能够识别出每个点所代表的那个人（或者个案）的一对分数，这是很重要的。设想这里已经有 6 个人参加了一个驾驶测试，其中第一个人只有 1 次练习机会，第二个人有 2 次练习机会，以此类推。有 4 次练习机会的这个人得了 105 分，在散点图上通过一个点表示出来，这个点在垂直方向的"练习"轴上对应的是 4 ，相应的，在水平方向的"分数"轴上对应的是 105。注意水平方向的轴通常叫做 X 轴，垂直方向的轴则叫做 Y 轴。



数据	
练习次数	测验分数
1	27
2	54
3	78
4	105
5	120
6	149

图 12.4 高度的正相关——练习的次数和驾驶测验的分数

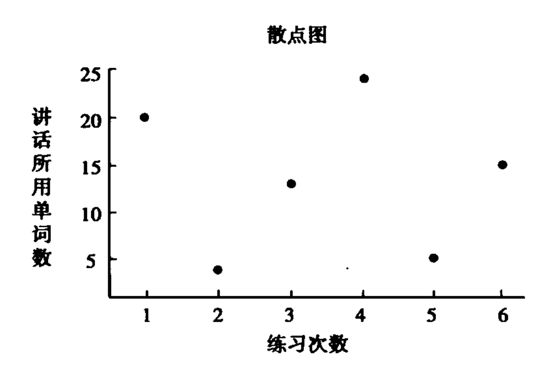
图 12.4 展示了一个非常强的正相关。然而，图 12.5 则展示了一个非常强的**负相关**——练习得越多，在任务中花费的时间就越少。注意这两个相关的形状，因为它们经常出现在考试中。图 12.6 中的这些点出现在每个地方，没有呈现出明显的线性形状。这就是我们所预期的两个变量之间没有相关。这里的相关系数的值很可能接近于 0。当两个变量的相关系数是零时，它们之间存在明显的关系的情况是罕见的。图 12.7 展示了一个**曲线关系**（curvilinear relationship）。该曲线代表所谓的唤醒（arousal）和操作水平两个变量间的关系：当我们厌倦（低刺激水平）时，我们



数 据

练习次数	完成路线的时间/秒
1	127
2	118
3	106
4	98
5	85
6	76

图 12.5 **高度的负相关——练习的次数和花在任务上的时间**



数 据

练习次数	讲话所用单词数
1	20
2	4
3	13
4	24
5	5
6	15

图 12.6 **零相关——练习的次数和任务期间讲话所用单词的数量**

以很低的水平完成；当我们接受适度刺激时，我们会完成得很好，但是当我们受到过度刺激时，可能会因为焦虑和负担过重而做得很糟糕。

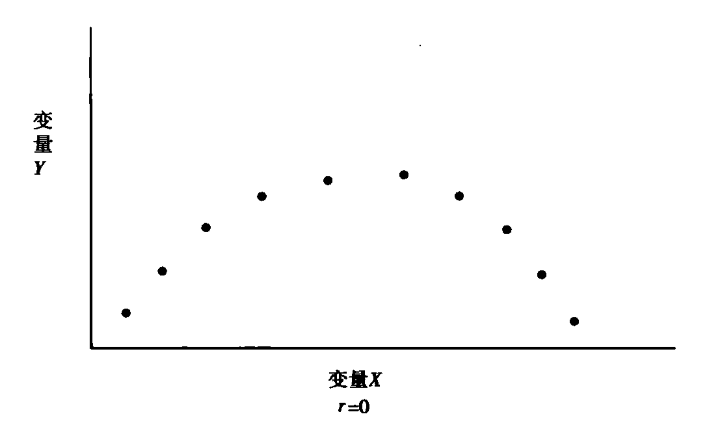


图 12.7 **两个变量之间的曲线关系**

红酒的消费量与什么呈负相关呢？

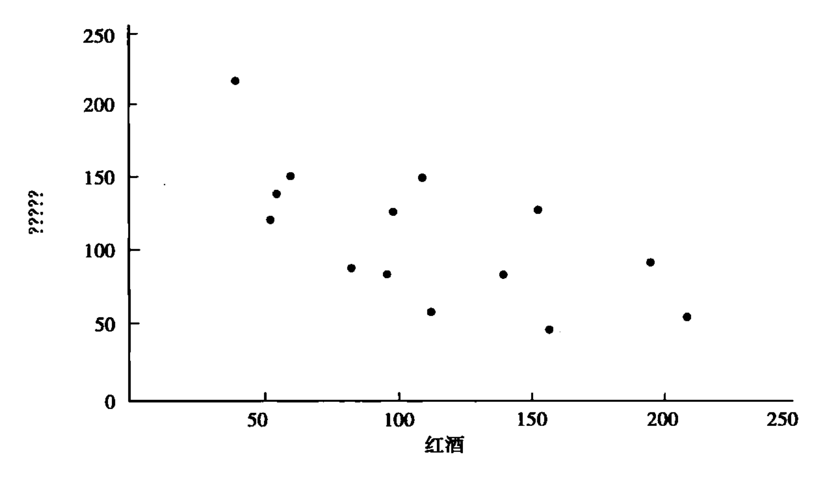


图 12.8 不同居住区红酒和神秘产品(*Y* 轴)的消费量
参照巩固练习 12.1 中第五个问题。
来源:Argyle(1994) *Psycholgy of Social Class*, London: Routledoe

练 习

1. 判断表 12.1 呈现的相关中,哪些是正相关,哪些是负相关。
2. 试着为表 12.2 中的数据画一个散点图。
3. 设计一些研究可以检验下面的目标:
   (a)熟能生巧。
   (b)吃鱼可以让人变得聪明。
   (c)权力使人腐败。
4. 一个同学告诉你,她得到的相关系数是 2.68。你会告诉她什么?
5. 乐一乐,请看一下图 12.8,然后试着猜一下在 *Y* 轴上的那个神秘变量。每个点都代表一个居住区域(例如,"富裕的郊区"(affluent suburban),"最差的贫民窟"(poorest council estate),等等)。在每个区域,红酒的平均消费量被记录下来,同样地,那种神秘产品的消费量也被记录下来。一些地区消费了很多红酒但却消费了很少的神秘产品,同时,另一些地区消费了很少的红酒但却消费了很多的神秘产品。它是什么呢?这些是一个真实研究里的数据并在图 12.8 的标题中被标注出来。
6. 你将如何描述在图 12.9 中呈现的相关的强度?

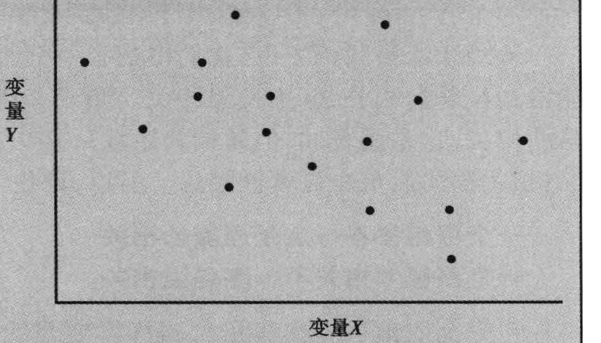


图 12.9 巩固练习 12.1 中问题 6 的散点图

答 案

1. ①正相关;②正相关;③负相关(年龄越大/记忆力就越差);④正相关;⑤负相关;⑥正相关。
2. 略。
3. (a)一个例子就是前面已经描述过的一个研究:给予被试或多或少的练习机会,在测试中记录他们的表现;

(b)我们需要一个量度标准来说明人们吃了多少鱼,如每月的重量,以及智力的一些测量尺度,如智商;

(c)得到他们所在组织中地位的等级水平(例如,经理、高级经理、总经理),用某种量表或设置一个使他们做出反应的情景,例如,对于他们来说,多少赔偿费是过量的?

4. 她需要重新计算。相关系数可能的最高值是 +1。

5. 浇汁。①

6. 中等程度的负相关。

关键词

相关(correlation)	负相关(negative correlation)
相关系数(correlation coefficient)	正相关(positive correlation)
曲线关系(curvilinear relationship)	散点图(scattergraph)

12.4 相关的显著性和强度

一个相关是强的与一个相关是显著的,这两种说法之间存在很大的差异,记住这点是非常重要的。

假设我有三张绿色奖券,依次编号为 1 ~ 3,三张红色奖券也被编号成 1 ~ 3。假设我随机给三个朋友中的每一个人:一张绿色奖券、一张红色奖券。每个朋友最后拿到两个相同号码并得到一个完全相关的概率是多少呢?如果我们持续进行这个令人烦躁的工作,我们预测完全相关发生的概率将是 1/6。只有 6 个可能事件,所以完全匹配发生的概率是 1/6 或者 0.17。如果这个虚无假设是真的,那么我们已经可以得到结果随机发生的概率。但是我们知道这种概率一定低于 0.05 的显著性水平。因此,在一个相关里只有三个被试是不可能得出一个显著性结果的。随着被试数目的增加,得到一个显著性相关的概率会变得越来越小。

从刚才给出的例子里,我们得到了一个非常强的但不是很显著的相关。我们也能得到显著的相关,但不是强相关。附录 2 中的表 8 给我们展示了一个有 60 人参与的相关,它是显著的,但其相关系数只有 0.25 那么低。科学家经常报告这样的情况:相关很弱但是有显著性趋势。所以,记住下面的内容是非常重要的:

一个强相关不一定是显著的相关。

一个显著的相关不一定是强相关。

它完全依赖于样本容量。当样本容量大时,一个非常弱的相关也可能是显著的。当样本容量小时,一个强相关甚至也可能是不显著的。这个消息是为了再一次确保你的样本里包含足够多的人。当你进行一个研究项目时,如果有密切的关系,你将有足够的人或者个案来说明它。

现在我们回到这个问题上:如何估计相关的强度以及怎样检查相关的显著性。

① 传统西餐不可缺少的调味料,一般不和红酒一起食用。——译者注

皮尔逊(积差)相关系数——r

什么时候使用皮尔逊相关系数 r

被检验关系的类型	数据的水平	设　计
相关	等距或等比数据	一组关系密切的成对数据

条件:满足参数假设,即数据必须至少是等距数据,而且应该符合正态分布。

优点:更加灵敏的检验——有时用斯皮尔曼相关得不到显著性的时候,用皮尔逊相关可以得到。

缺点:需要满足参数假设;计算复杂。

这个令人印象深刻的话题,包括“积差”,不经常使用。这个检验经常被简单地叫做皮尔逊相关。然而,复杂的题目是与相当复杂的计算方法相匹配的。你用斯皮尔曼等级相关(在后面将会描述)和皮尔逊相关可以做出差不多同样的结果,但是后者是一个更加灵敏的检验。这反映出 t 检验和在后面一章将提到的威尔科克逊检验(秩和检验)之间的关系。偶尔,我们能用皮尔逊相关得到显著但是用斯皮尔曼相关得不到显著,可参照我们后面提到的计算方法。每种方法均有其各自的优缺点。

我们将对表 12.2 的数据进行皮尔逊相关的计算。在“自尊”和“自信”这两列里的数据是原始数据。假设我们就这两个变量对这 10 个被试进行了施测;尽管自尊量表的最高分是 20 分,其自信量表的最高分却可能是 30 分。注意这两个相关的变量中的每一个必须能够用等距或者是等比水平的量表来测量;他们不必有相同的全距。如果我们愿意,我们可以计算温度和年薪之间的相关程度。我们应采用的公式是:

$$r = \frac{N\sum(xy) - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

我担心这个公式看起来相当复杂。这里有几个关于这个等式的不同版本,所以不用担心它看起来不同于你已经在别处看到的那个等式。这个公式的优点是你不必计算平均数(mean)或者标准差(standard deviation),但是你能够根据表 12.2 最后两行给出的值用一个简单的计算器计算出来。

稍作解释

皮尔逊相关是通过观察**离差分数**(deviation)得到的。这就意味着,如果 x 和 y 具有强相关,那么就会存在这样的趋势:一个人在一个变量上的得分高于平均数,那么他在另一个变量上的得分也将高于平均数,反之亦然。因此,他们的离差分数都高。在皮尔逊相关的计算里,离差分数是相乘得到的(尽管上面的公式中并没有直接体现出这一点)。当两个离差分数的变化趋势完全相同时,我们得到了所能得到的最大值。例如,如果我在 x 上的平均分是 5,在 y 上的平均分是 5,那么我们将会得到 25,然而,如果我在 y 上的平均分只有 2,那么我们只能得到 10。库利坎(Coolican,2004)对皮尔逊相关计算方法进行了更加全面的解释。

表 12.2　10 个人的自尊和自信分数

被试	自尊 x	自信 y	x^2	y^2	$x \times y$
A	17	25	289	625	425
B	8	12	64	144	96
C	3	8	9	64	24
D	12	8	144	64	96
E	11	10	121	100	110
F	18	22	324	484	396
G	6	19	36	361	114
H	10	14	100	196	140
I	11	16	121	256	176
J	14	23	196	529	322
	$\sum x = 110$	$\sum y = 157$	$\sum x^2 = 1\,404$	$\sum y^2 = 2\,823$	$\sum (x \times y) = 1\,899$
	$(\sum x)^2 = 12\,100$	$(\sum y)^2 = 24\,649$			

皮尔逊相关系数的计算

过　程	计算步骤/结果
1. 找到 $\sum x$;	看表 12.2 的第二列等于 110
2. 找到 $(\sum x)^2$（这是 $\sum x$ 本身的平方）;	看表 12.2 的第二列等于 12 100
3. 将所有的 x^2 相加得到 $\sum x^2$，仔细区分 $\sum x^2$ 和 $(\sum x)^2$;	看表 12.2 的第四列等于 1 404
4. 将第 3 步的结果与 N 相乘;	$1\,404 \times 10 = 14\,040$
5. 从第 4 步的结果中减去第 2 步的结果;	$14\,040 - 12\,100 = 1\,940$
6 ~ 10. 在 y 的数据上重复第 1 ~ 5 步的过程;	从表 12.2 $\sum y = 157$ $(\sum y)^2 = 24\,649$ $\sum y^2 = 2\,823$ $N\sum y^2 = 28\,230$ $N\sum y^2 - (\sum y)^2 = 28\,230 - 24\,649 = 3\,581$
11. 将第 5 步与第 10 步的结果相乘;	$1\,940 \times 3\,581 = 6\,947\,140$
12. 找到第 11 步结果的平方根;	$\sqrt{6\,947\,140} = 2\,635.74$
13. 将 $\sum x$ 与 $\sum y$ 相乘;	$110 \times 157 = 17\,270$
14. 找到 $\sum xy$;	看表 12.2 的第六列，$\sum xy = 1\,899$
15. 将第 14 步的结果与 N 相乘;	$10 \times 1\,899 = 18\,990$
16. 从第 15 步的结果中减去第 13 步的结果;	$18\,990 - 17\,270 = 1\,720$
17. 用第 16 步的结果除以第 12 步的结果。	$r = \frac{1\,720}{2\,635.74} = 0.653$

为了查明 r 所得的这个结果是否是显著的，请用 $df = N - 2$，参照附录 2 中的表 8。单侧检验适用于检验正相关或负相关。仅进行一个方向的检验意味着另一个方向的检验在逻辑上没有任何心理意义，这显然是不可能的，因此，我们几乎总是进行双侧检验。这里，$df = 8$；$p \leqslant 0.05$ 的临界值是 0.632（双侧）。我们得到的值刚好大于这个值，因此是显著的。

如果 N 是大于 100 的，那么 r 就能被转化成 t 值，然后在 t 值表中进行检查。

详细描述相关的结果

自尊量表和自信量表上得分的皮尔逊相关是显著的，$r(8) = 0.653, p < 0.05$，双侧检验。

负相关

如果你的 r 值正像你预测的那样得出是负的，这是可以的——看一下这一章开头有关负相关的解释。对显著性检验来说，我们只考虑 r 的取值不考虑它的负号。

斯皮尔曼等级相关系数(ρ)

什么时候用斯皮尔曼相关系数 ρ

被检验关系的类型	数据的水平	设　计
相关	等级数据	一组关系密切的成对数据

优点：不需要满足参数假设。更容易计算。

缺点：与皮尔逊相关相比，对数据的灵敏度较差，因此当用皮尔逊相关得出显著时，用斯皮尔曼相关不一定显著。

对这个相关的计算方法，我们能够用和上次一样相同的数据，然后对结果作一个比较。斯皮尔曼的计算公式是：

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

其中 N 是这列数据的数目。d 是每一对分数排列等级的差数，就像下面解释的那样。

略作解释

正如威尔科克逊检验一样，我们看一下成对等级序列之间的差异。然而，如果在这里有一个很强的正相关，那么我们将期望所有等级之间的差异是小的。每个人将在两个变量上得到几乎相同的等级。如果我们想得到一个很强的负相关，那么人们应该在一个变量上得到很高的等级，而在另一个变量上得到很低的等级，反之亦然。在斯皮尔曼相关的计算公式里，如果差异很小，那么 $6\sum d^2$ 也会很小，然后整个分数也将是小的。因此，只能从 1 里减去很小的一部分，结果 ρ 将是接近 1 的。斯皮尔曼相关实际上是一个皮尔逊相关，但是，前者是用每个分数的等级算出来的，而不是用分数本身，正如皮尔逊相关表 12.3 的第四和第五列（从左看）所表示的

那样。

表 12.3 计算斯皮尔曼相关所用到的自尊和自信两个量表的分数

被试	自尊 x	自信 y	自尊等级	自信等级	等级之间差异 (d)	d^2
A	17	25	9	10	-1	1
B	8	12	3	4	-1	1
C	3	8	1	1.5	-0.5	0.25
D	12	8	7	1.5	5.5	30.25
E	11	10	5.5	3	2.5	6.25
F	18	22	10	8	2	4
G	6	19	2	7	-5	25
H	10	14	4	5	-1	1
I	11	16	5.5	6	-0.5	0.25
J	14	23	8	9	-1	1
						$\sum d^2=70$

斯皮尔曼相关系数的计算

过程	计算步骤/结果
1. 给 x 的分数一个等级；	看表 12.3 的第四列
2. 给 y 的分数一个等级；	看表 12.3 的第五列
3. 用每一个 x 的等级减去每一个 y 的等级；	看表 12.3 的第六列
4. 将第 3 步的结果平方；	看表 12.3 的第七列
5. 将第 4 步的结果相加；	=70(看看表 12.3 中第七列的下面)
6. 将结果代入到前面给出的斯皮尔曼相关的计算公式里：	$\rho=1-\frac{6\times70}{10(10^2-1)}=1-\frac{420}{990}=0.576$
7. 将结果与附录 2 表 9 中的临界值相比较。	$N=10$。用双侧检验，要达到显著，ρ 的值必须比 0.648 大。在这个案例中，我们得到的相关是不显著的，所以必须接受虚无假设：没有相关。

注意这里，我们用皮尔逊相关得出了显著性，但是，用斯皮尔曼相关却得不到。这就是参数检验与非参数检验在计算时检测力上的差距。

如果样本量比较大，即 N > 30

在对相关系数的显著性检验中，你可能已经看出在斯皮尔曼相关的表格中，N 最大只能到 30，在皮尔逊相关中 N 最高可以到 100，但是，在一个调查里，经常会遇到多于 100 人的时候。那我们应该怎么做呢？方法是转换你的 r 值，然后像我们在上一章做的那样，检查 t 的显著性。转化的公式是：

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

它同时适用于皮尔逊和斯皮尔曼两种相关。

12.5 相关的一些概括

相关并不意味着因果关系

回想一下在前面提到的那个冷饮的例子。我们在饮料的销售额和温度之间做了一个相关。如果我说"好的,那个相关是卖饮料让温度升高的证据。"那么你将认为我非常奇怪。因为你知道销售额是不可能影响温度的。温度升高是饮料销售额增加的一个直接原因,这是很有可能的。那是因为你从生活的常识里很清楚地知道什么导致了什么,比如上述那种情况。然而,在心理学里,没有什么事情是那样确定的。思考一下下面这个相关的表述:

孩子受到的体罚越多,他们的攻击性就越强。

体罚和后来的攻击性之间的相关已经在很多研究中都有所描述。然而,上面的结论却很难解释清楚。这个结论很容易使我们作出这样的推断:体罚儿童**导致**他们更具攻击性。但是,仅以相关为依据,我们只能说二者之间存在一定的关系。天生具有攻击性的孩子很可能会得到来自父母的更多的体罚!我个人对这种观点表示质疑,但是,在科学研究中,在得到更直接的证据之前,我们必须保持一个毫无偏见的头脑。在这个案例里,我们从一个自然实验里得到了更加直接的证据。1979 年,在瑞士,包括父母在内的任何人打骂孩子是不合法的。有人预测,缺乏教养的充满暴力的孩子在街上游荡造成很大的破坏,但恰恰相反,达兰(Durrant,2000)发现,这一阶段,该因素①对这一社会现象②的预测力甚至强于其他(影响其他相似社会)因素的预测力。

相关的解释

当我们知道在 A 和 B 之间有一个相关时,就有几种可能的解释,包括以下几种表达:

- ❑ 变量 A 对变量 B 有一个直接的影响。
- ❑ 变量 B 对变量 A 有一个直接的影响。
- ❑ 变量 A 和变量 B 都受变量 C 或者其他某个变量的影响。
- ❑ 这个相关是个 I 型错误——这只是侥幸。

上面提到的第三种选择是说变量 A 和变量 B 都被另外一个变量影响。例如,攻击性可能不是被体罚引发出来的,体罚也可能不是由孩子的攻击性所导致的。体罚

---

① 指 1979 年有关打骂孩子的法律的实施。——译者注

② 指青少年犯罪率下降。——译者注

和攻击性这两个变量可能都跟家庭环境的某种形式有关系。那些体罚孩子的父母可能在做其他事情的时候鼓励攻击性，例如，他们自己就具有攻击性或者没有抑制住攻击性行为。这些可能的解释在图 12.10 中被描述出来。

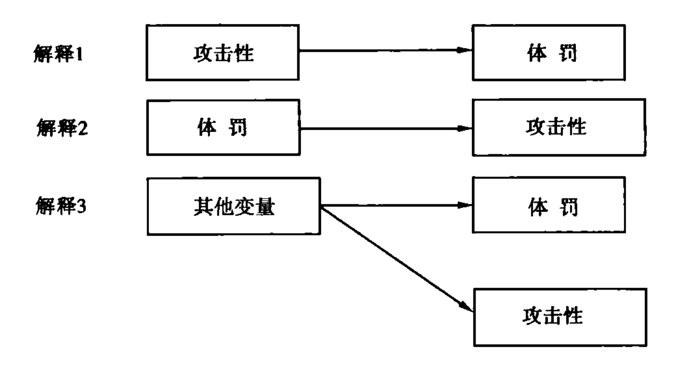


图 12.10 在体罚和攻击性之间相关的可能的解释

相关和实验研究的比较对照

就像刚才我们看到的那样，相关研究的**不足之处**在于**单独的**相关不足以告诉我们哪一个变量在哪一个方面产生了影响。因为控制很好的实验会操纵一个自变量(*IV*)，然后观察其在因变量(*DV*)上产生的影响。我们能更有自信地说是 *IV* 引起了 *DV* 的改变。相关使我们只能用两个变量间关系的存在作为一个更为间接的证据，来说明一个变量对另一个变量有影响。

相关研究的**优点**在于当用于比较的实验是不可能实现的时，它们经常可以被采用。我们不可能要求人们抽很多的烟，然后看他们病得有多厉害。然而，我们可以询问他们抽了多少烟，然后将这个变量与其他变量做一个相关，例如生病的天数，看医生的频率以及与吸烟相关的疾病发生的次数，等等。

确保你的变量都是可测量的

在一个相关里，**必须**确保两个变量至少是顺序数据。它们几乎总是"分数"，看起来像等距数据。当选择斯皮尔曼相关后，这些分数可以或者不用转化成等级数据。当你获得每个人的一个得分和一个分类变量时，你不能使用相关。例如，你**不能**把一个人的性别与其外向性的得分进行相关。性别属于一个分类变量。这里很理智的做法就是在有男女两个水平的性别(rank)这个"自变量"(independent variable)上做一个**差异**检验(difference test)。我们将寻找男女在外向性(extroversion)上的差异性。实际上，刚才说不能对这里的性别做相关是一个特殊情况。如果这个分类变量(categorical variable)只包含两个值，例如男/女或者自家车/非自家车，那么你可以对它做相关。然而，这些过程对你的课程提纲来说可能太超前了，但是，你可以在库利坎(Coolican,2004)的著作中找到你想要的。这里有一些学生经常试图用之做相关，并带来非常严重后果的典型的分类变量：

- 婚姻状况(如果你把"1"看做单身，"2"看做已婚，那么结过两次婚的单身怎么去赋值呢?)
- 地区(例如北方/南方)

- ❑ 宗教信仰
- ❑ 私家车的类型
- ❑ 左撇子/右撇子(left-handed/right-handed)
- ❑ 可存放的/不可存放的
- ❑ 音乐偏好

练 习

1. 1988 年 6 月 3 号出版的《泰晤士报》教育副刊中的一篇文章:

   "…… 把字母的声音和形状教给学前儿童可能是一个好的开始。4 岁时受到良好基础教育,读写算能力良好的孩子,7 岁时倾向于表现得好。

   在阅读这件事情上,这个孩子在 3~4 岁的时候所能够识别的字母的数目是预测 7 岁孩子能力的最强因素。蒂泽德(Tizard)推断托儿所的老师应该多强调识字和算数的能力。"

   (a)除了研究者推断出来的,我们还能从文章的最后一段推断出什么结论呢?

   (b)简单地描述一个能够帮助我们作决定的研究。

   (c)研究者已经发现的,4 岁儿童识别字母的数目与其 7 岁阅读发生错误的次数之间的相关是哪一种类型的相关?正相关还是负相关?

   (d)假设孩子在 5 岁时的加法能力与其在 7 岁时的数学能力之间的相关系数是 0.83。你怎么用言语描述这个相关系数的强度?

   (e)如果样本里有 33 个孩子,那么这个 0.83 的(皮尔逊)相关达到(双侧)了什么样的显著性水平?

2. 斯皮尔曼相关总是被用来替代皮尔逊相关。这个论断反过来说对吗?请给出原因。
3. 一个研究者把被试在一个和自尊强度有关的问卷上的得分,与通过评估他们对几幅图片的言语反应获得的焦虑水平之间做了一个相关。这个相关应该采用哪一种计算方法呢?
4. 一个朋友想要做一个练习。在这个研究里,她问人们是去官方学校(state school)、私立学校(private school)、公立学校(public school)还是其他类型的学校。她还要求他们填一下有关学习态度的问卷。现在她想把教育类型和学习态度之间做一个相关。你建议她做什么?

答 案

1. (a)早期 4 岁儿童的字母识别能力与其 7 岁的阅读能力相关,但是不太可能**导致**出众的阅读能力。可能是一般的家庭环境影响了早期的字母识别能力和 7 岁时的阅读能力。加强字母的认知能力可能不会自动导致 7 岁时阅读能力的增加。

   (b)你可以设计一个 4 岁的实验组(experimental group),与一个控制组(control group)相匹配,训练实验组儿童的字母识别能力。紧接着是 7 岁的那一组,然后比较各组在阅读能力上的差别。你也能平衡掉你可以得到的所有的第三个因素(尽可能地思考引起 4 岁儿童识字能力和 7 岁儿童阅读能力提高的原因),然后看看从 4 岁到 7 岁的各个组的阅读能力。

   (c)负相关;技能越好,产生的错误越少。

   (d)强/非常强。

   (e)$p \leq 0.001$。

2. 不对。因为对皮尔逊相关来说,数据必须满足参数的要求。
3. 斯皮尔曼相关。因为评估的是人类的判断能力,所以数据应该被看做顺序数据,而且这种测量也不能被描述成标准的测验。

4. 她的教育变量的类型属于分类变量，所以不可能做出相关。她能：

(a) 把她的态度数据分解成两组（例如，在中位数以上和以下），并且把学校种类也分成两组（例如，官方/私立学校），然后做个卡方检验。

(b) 仅仅把学校类型分解然后进行一个非相关的差异性 t 检验。

关键词

皮尔逊积差相关（Pearson's product-moment correlation）

斯皮尔曼等级相关（Spearman's rho(ρ) correlation）

13

心理学研究中的道德问题

本章内容

- ❑ 心理学家在一些领域的表现必须符合道德标准，包括接待来访者，出版自己的著作，以及进行一些人类被试参与的学术研究。本章对于上述前两个方面做了简单的介绍并集中探讨了第三个方面的问题。
- ❑ 了解英国心理学会关于以人类为研究被试的道德原则。
- ❑ 讨论保密性与匿名性的问题，这两部分的内容经常被混淆。此外，回顾了关于隐私的问题。
- ❑ 重新讨论对于欺骗的态度问题并提出可能的替代方法。
- ❑ 描述了对于任务报告的必然需求，并针对任务报告是否有效提出问题。
- ❑ 确保被试心理和身体得到保护是最重要的，围绕这个问题展开包括被试的心理压力在内的各种各样的争论。
- ❑ 参与调查前，研究者必须让被试了解并同意所参与研究项目的程序。然而，如果告知的信息中存有欺骗的成分，那么就不能一再地说被试是完全知情的。我们将把这种暗示跟调查者的特殊权利一起讨论。
- ❑ 在一定距离外观察人们意味着被观察的对象是没有征得同意的被试。因此这就涉及了无意识参与的问题。我们将会列举一些程度严重和轻微的案例。
- ❑ 最后，我们将处理一些当心理学者进行干预性研究时会涉及的道德问题。干预性研究会对被试的生活产生明显的影响。然而这种关于描述过的心理效应的有益影响并不可能推广到所有人，例如阅读能力或是认知能力的提高。

作为专门处理公众事务的专家，心理学者在工作时必须遵守一定的道德标准。因为作为研究者或是从业者。他们对工作对象的日常生活有某种特定的权力和影响力，因此研究必须要保护公众，这样才能让公众信任他们遇见的心理学者，并且相信心理学者不会暴露他们的信息或对他们有害。

事实上，我们当然都同意，不管是故意的还是无意的，心理学者们完全不应该伤害或暴露被试。主要有三个领域中的道德问题很重要：

1. 在工作的应用领域里，心理学者应该作为专业人员来帮助普通大众。
2. 心理学家发表的心理学研究成果可能会被媒体和广播无意中传播给公众。
3. 心理学家与参与研究的普通大众一起工作。

13.1 作为从业者的心理学家

许多心理学家作为从业人员在工作。这就是说，他们运用心理学知识和专门的技术尝试在日常生活中帮助人们。他们也许是教育心理学家、运动心理学家、治疗学家、咨询师或是其他的一些专家。这些心理学家同来访者一起工作，而且对于这些来访者来说，心理学者有着职业操守。

在美国，有时候一些来访者会向美国心理学会的道德委员会投诉，然后由道德委员会进行裁决。美国心理学会的标准是有强制性的。被投诉的心理学家会受到严责，被心理学会开除，或是被要求改变他们的行为或接受相关的培训。这种原则适用的广度和惩戒的力度反映了美国心理学会对于作为消费者的公众有着良好的适用性。

自从 1987 年以来，英国皇家心理学会的宪章已经得到修改，因此我们的一些方式向着上文提到过的美国模式转变。现在学会支持特许的心理学者的注册。他们是在应用领域或是研究领域实践心理学的人。这些登记的成员使用标准的字母"C. Psychol"，他们会因为不专业的行为而被除名，并且大家都希望，他们能成为被人们所认可的有诚信的"有标志的"从业者，从而被公众所认知和信赖。

1993 年，英国心理学会发布了《操作原则》并且希望所有向学会登记过的心理学家遵守《操作原则》(参见英国心理学会制作的《操作原则》,2000)。随着这本书的出版，于 2006 年重新探讨这些原则并且采纳了修订的原则。然而，我们更倾向于讨论研究方法而不是应用的实践，因此我们也不花费更多的时间在这些操作的原则上了，但是为了其他人的兴趣，在下面的网站可以看到更多关于这个方面的内容：

http://www. bps. org. uk/the-society/ethics-rules-charter-code-of-conduct/code-of-conduct/acode-of-conduct-for-psychologists. cfm

13.2 心理学研究成果的发表

作为整个研究的一部分，心理学者有责任去发表那些有良好的并有普通理论支持的文章。这些结果和研究方法也必须能够被那些希望去分析、证明并且/或是在某种程度上复制这些结果的感兴趣的研究者利用。研究者必须重视研究结果可能

引起的社会影响并且评估这些结果在主流社会中,对于道德和政治思潮的价值。心理学研究领域中,20 世纪最重要的并备受争议的课题也许是假定"人种"在智力上是有差别的。一个"人种"的平均智力水平明显高于另一个"人种",这个没有根据的观念曾经激起了校园暴乱。令人沮丧的是,这个观念还曾经被极端的政治集团利用,用来论证应对于不同的种族群体实施不同的教育方式,并消除有关教育和福利计划的支持。"人种不同"论的支持者认为,如果这个所谓的"人种的"智力水平不同的原因在很大程度是由基因决定的,那么这些计划是没有作用的。支持这种差异的证据是极其微弱的,并且这样的情况通常很少被认同。再者,这个问题对于本书的目的太过于具体和复杂,然而在库利肯(Coolican)2004 年作品的第 20 章或是在例如理查兹(Richards,1997)的专门性文章中你可以找到更多的内容。此处我们只是简单地陈述一个结果,发表文章的心理学者们有一个公开的职责,即考虑他们经常进行的试验性质的发现对于公众可能产生的影响和反馈,因为他们会无意识地产生有关心理学研究临时性的和有限的暗示。

英国心理学会关于人类研究的道德所做的规定

2000 年,英国心理学会发布了一个手册。手册内容覆盖了极其广泛的道德话题,包括早先提到的《操作原则》和 1992 年被学会采纳的《以人为研究被试的道德原则》。这些原则可以在下面的链接看到:

http://www.bps.org.uk/the-society/ethics-rules-charter-code-of-conduct/code-of-conduct/ethical-principles-for-conducting-research-with-human-participants.cfm

作为操作一个小型研究课题的学生,最好在开始前就谨慎地查阅这些原则。你应该把这些原则作为最低限度的一套指南去遵从;你也应当接受导师关于你计划进行的课题的建议:对你而言课题中什么是有可能,哪些是有道德危险的。英国心理学会建议成员应遵守如下内容:

成员也应该标出这些原则而引起非学会成员的研究同事的注意。成员应该鼓励同事去采纳这些原则,并确保所有他们指导的研究者来遵循这些原则(例如,研究助理、研究生、本科生、普通教育文凭和普通中学文凭的学生)。

这些原则从总章开始,强调被试的利益和感受,以及研究者从被试总体的角度检查研究背景的需要。研究者不能臆断自己已经很好地了解了什么样的行为是被试能够接受的,什么样的行为是不能接受的。

在所有的情境中,调查者必须考虑到对于研究被试所产生的道德上的意义和心理学的后果。基本原则是调查应该从所有被试的角度考虑;应该消除对他们心理幸福、健康、价值和尊严会产生的可预见的威胁。调查者应该认识到,在我们这个多元文化和多元道德的社会中,调查会涉及的不同年纪、不同性别和不同社会背景的个体,调查者对于影响被试的调查也许并没有足够的知识。但应该铭记的是,一项调查的最佳判断是否会冒犯这个群体中的成员,导致这些成员在该项研究中会被贴上标签。

英国心理学会道德研究的原则主标题列举在专栏13.1中。我们应该参照这些原则，同时讨论对于人类被试的各种伦理问题的研究。应该指出，在今天英国的大学中一直有一个道德委员会，对于我们可能考虑到的所有领域的研究计划（人员和学生的）中那些无法被接受的程序进行再次探讨。

专栏13.1　英国心理学会关于人类被试的研究操作的伦理原则的主标题

1. 同意（包括调查者对被试影响力的常见问题）。
2. 欺骗（那些会降低公众对心理学研究的信任的欺骗）。
3. 任务报告（在调查实施后完全告知被试调查研究的内容并在实验后让被试能够恢复到实验前的状态）。
4. 从调查研究中退出。
5. 机密性（对于结果和产生结果的人的保密；这包括大量匿名的使用）。
6. 对于被试的保护（避免对他们身体和心理产生危害）。
7. 观察研究（包括在公共场合观察的问题）。
8. 给被试建议。
9. 同事（确保这些同事也能遵守道德标准进行工作）。

13.3　对人类被试的研究操作

机密性、匿名和隐私

除了所有道德上的考虑，始终向被试担保匿名性是个纯粹务实的看法。如果心理学家把结果连同被试身份一起发表，公众会马上停止成为志愿者或是拒绝同意参与研究。

一个调查能确保匿名性，通过向被试保证在发表数据的时候，永远都不会泄露每个被试的身份。这件事必须经过深思熟虑，如果一个人的细节被透漏，即使没有说出名字，他们的身份也会被确认（例如，一个在小公司进行的应用型的课题）。调查者也可以要求获得暴露个体身份的允许。这样的暴露也许只有当使用影像记录时出现，例如在米尔格拉姆服从实验中的记录（Milgram，1974）。那些已经被严重欺骗的研究被试可以运用所有被试的基本权利，亲眼目睹销毁他们不希望保存的任何记录。如果要保存记录，被试有权利确认他们的个人信息是安全的，而且只有那些被完全信任的研究工作人员才能将这些记录作为匿名数据使用。

这里我们需要区分一下匿名性和机密性的概念。如果发表结果（例如访谈），那么这些不是机密的，但是它们仍然可以保持匿名性。事实上，一些研究的数据对于其他研究者也是可以使用的，而保持这些数据的机密性对研究的结果没有多大用处。然而，机密更多的是一种心理学实践的特性（保持来访者数据的机密）而不是研究和发表的特点。

有些特殊的情况，调查者也许要违背匿名性的规则，就是当人们的生命有明显危险的时候。一个调查者对有危害的生活进行参与观察，有一个清楚的责任就是当

有严重的犯罪行为发生时候，就不能保持机密性了。一个精神病治疗的病人打算杀死他自己或是室友，这就要报告。在这里有关的道德原则范围就比在科学研究操作过程中的宽泛些。

被试显然有保护隐私的权利，并且实验过程不应该没有提前说明，或是没有告诫地就直接侵犯被试的隐私。实验过程中如果有潜在的、私人的、令人局促不安或是敏感的环节，必须清楚地提醒被试，他们有权利保留信息或是拒绝参与。有些需要特别注意的地方，例如，在询问被试关于性态度或性行为的时候。

在暗中观察研究的情况下，这个原则遵从起来很困难，并且在这样的情况下，不遵从规则的使用者也会使严肃的批评升级——以下看汉弗莱斯（Humphreys，1970）的例子。这样的情况下很难检测，报告的最后版本是为了证明他们对于被试的发言内容做了精确的叙述。

对于米尔格拉姆 经典实验的道德讨论

在心理学研究中，任何道德原则的讨论都无法避免地要对早期米尔格拉姆著名的服从实验程序的讨论。在这个研究中，涉及几个道德问题，所以让我简短地描述一下，然后你去思考下这些问题是什么。

专栏 13.2 对于米尔格拉姆服从范例的简短描述

实验部分需要和大家讨论之后再做修改。“被试”，而这些“被试”实际上是实验助理。志愿者成为实施电击的“老师”，对于助理产生的每次错误都提高 15 伏特的电压。375 伏特被认为是“危险：剧烈的电击”。一个磁带录制的尖叫和拒绝，欺骗了老师-被试，并使他们相信实验助理在经受巨大的痛苦并且希望结束这个部分。通过实验者的敦促，例如“实验需要你继续”和“你除了继续进行别无选择”，而迫使老师-被试继续进行。令米尔格拉姆诧异的是，即使在 315 伏助理已经停止反应，65% 的被试在一定底线的范围内（450 伏）仍会继续释放电击。米尔格拉姆同“富有经验但客观中立的同事”商议，他们预期不超过 1% 的被试会服从到最后。老师-被试经常会显示出极大的焦虑。一个被试甚至出现了轻度的痉挛。一个观察者写到：

> 我看到一个最初泰然自若的成熟的商人带着自信的微笑进入实验室。20 分钟后，他已经处于一种颤动的状态。口吃非常严重，已经很快就要接近于紧张崩溃的临界点。他不断地拉扯自己的耳垂和扭曲自己的手。在这时，他用拳头捶击自己的前额，并喃喃自语：“哦，上帝啊，让我们停止这些。”（Milgram，1974）

这个实证的结果可以用来证明，许多普通人在压力下都会以一种相当残酷的方式来行动。暴行并不必然只会被完全邪恶的人或是“文化”所实施。

列举一下这个调查中你认为不符合道德的方面。这样的调查到底该不该实施？可以根据结果（科学的和令人惊奇的知识）而判断调查的意义吗？

欺 骗

米尔格拉姆的被试被严重欺骗了。研究者不仅令他们相信他们正在电击一个无辜的受害者,受害者正遭受着巨大痛苦,而且呈现给被试的是惩罚影响学习的实验,整个研究的目的完全被扭曲。

欺骗或者至少隐藏信息的行为在心理学实验中是极其普遍的。在一个权威研究杂志《人格与社会心理》上的一篇调查文章显示,研究中的欺骗率在20世纪90年代与70年代是一样高的,约为47%(Sieber,Iannuzzo and Rodriguez,1995)。对这个数据还存在争论,一些人认为要低些。因为对于"欺骗"的定义不同,在百分比上也可能不同。47%的数据包括在研究中不告诉被试实验的真实目的,然而如果对欺骗的定义严格些,明确地说"告诉被试不是真实的事情"(Ortmann and Hertwig,1998),那么数据会变得低些。一些人(Ortmann and Hertwig,1997)认为,心理学家应该按照美国心理学会所给予的道德标准,支持在调查中完全禁止欺骗的使用。另外一些人认为这样的禁令也许会使心理学的研究变得无效(Kimmel,1998)。

一些欺骗看起来是非常明显没有害处的。实验者告诉一些被试一个婴儿是个男孩,对另一些被试说成是女孩,然后比较他们对于婴儿的描述。被试执行的是一个感觉运动的任务,并告诉他们,观察者在现场是为了记录他们有关的技能性行为的细节,而实验真正的目的是记录观察者对他们执行任务的影响。告诉孩子们不能够玩玩具,因为这个玩具属于另一个在隔壁的小孩。告诉学生他们实验的小老鼠是"机灵的"。甚至是安慰剂的使用也可能是欺骗,虽然经常告诉被试也许可能使用了一种安慰剂。

一些欺骗可能严重些。告诉被试测试的结果显示他们没有好好地调整。对女性被试反馈说,认为她们有吸引力的或是没有吸引力的男性会在不久后拜访她们。布拉梅欧(Bramel,1962)给一些男性被试关于他们对于男性照片的情绪反应以错误的反馈,例如他们的反应看起来跟同性恋有密切关系。在拉托尼和达利(Latané and Darley,1976)的实验中被试认为他们无意中听到了真实的癫痫发作。因变量就是报告发作的速度或是频率。

如果使用欺骗,那么调查者可以做些什么呢?第一,英国心理学会1992年的原则建议,无论什么可能的情况下,对于个人的咨询操作必须结合来访者的社会和文化的背景。第二,任务报告应该小心地处理。第三,在一些事例中,获得欺骗的允许也是可能的。可以询问志愿者他们准备选择参与到哪种研究中,例如:

- ❑ 对商品认知的研究。
- ❑ 对产品安全的研究。
- ❑ 在这之前被误导的意图的研究。

任务报告

在被试正式参与实验之前,要让他简要了解情况,即告诉他们研究调查的整体情况(如果要使用欺骗的手段就不要让他得知实验的真实目的)并且对要求他们做的事情进行说明。不要将事后的和事前的描述混淆在一起。在所有的调查研究中,

研究者有责任向每个被试解释任务报告。在实验完全结束时要揭露研究的真正意图和目标,并且每次所做的尝试都要确保被试完成任务离开时,他们的自身感受跟他们刚到的时候是一致的。作为一个例子,约翰斯顿和戴维(Johnston and Davey, 1997)进行了一个实验,实验中的一个情景就是让被试听只有几条消极新闻的磁带。因为这也许会扰乱情绪,所以在告诉被试关于研究的所有情况和给他们费用之前,研究者让他们在个人的录音室听了两分钟的放松磁带。

如果研究涉及严重的欺骗,任务报告的责任还包括对于再次保证和解释的大量努力。这个任务报告本身也许不得不涉及一些欺骗,就像不管孩子们事实上表现的标准怎么样,都会告诉他们"做得非常好"。在米尔格拉姆的实验中,那些施加了极限电压的被试都会被告知,一些人参与这个研究感觉"非常愉快"。这个做法是为了帮助顺从的被试把他们自己不愿意进行和经历的焦虑,同编造出来的愉快顺从的被试的非常愉悦作比较(米尔格拉姆从没有报道过任何被试进行得很愉快)。即使是有这样一个对比,当他们离开时,40 人中至少有 26 人认为,他们在压力下可以对一个无辜的人类施加巨大的痛苦,只要不至死亡。让这些人在离开实验室时对他们自己的感觉跟他们进入实验室前一样,这几乎是不太可能的。

任务报告是否奏效

米尔格拉姆让他的被试在研究后完成了一份调查问卷。84%的人表示,能够参与实验他们很高兴,然而仅有 1%的人表示很后悔参与实验,其余人报告了中性感觉。80%的人认为应该有更多的类似于米尔格拉姆的实验研究进行。75%的人认为实验很有意义和对自我有启迪。一些研究者认为,不能完全接受米尔格拉姆研究的道德可接受性的判断,以及对这类研究的广泛尝试的赞赏。林、沃尔斯顿和科里(Ring, Wallston and Corey, 1970)决定评估这个研究对被试的影响,结果显示即使是研究者对米尔格拉姆的实验也心存不满,不仅对实验中的欺骗行为不满,还对先使用不真实的任务报告,后来才改用真实报告的做法不满。他们表示,最初的表面的任务报告极大地减少了被试去评估研究消极性的倾向。然而,他们也发现 1/3 的被试反应,即使在第二次完全的描述后,他们还是会对自身有一点的生气和失望。事实上即使只有少数的被试在实验后会感到对自己十分不利,还有许多被试在实验中就感到极其不安,许多研究者认为在道德上是无法接受这些如此严重的欺骗和压力的。

除了道德上的问题,调查者牵扯大量的欺骗也是不明智的。学生们会经常怀疑他们参加的研究的明显的结果和解释都是假的。林(Ring, 1970)以及其他人发现,50%的被试声称他们对于以后的心理实验会更加小心和怀疑。许多研究者显示,被试们表示不介意被欺骗。然而,泰勒和谢泼德(Taylor and Shepperd, 1996)研究显示,即使实验是真实的,被试的行为还是受欺骗常识的影响。他们安排被试了解到在实验的过程中欺骗了他们。这些被试就像什么也没有发生一样继续参与而且他们也不会揭露他们知道谎言的事实,即使试验者多次询问他们在实验过程中是否注意到什么奇怪的或是令人怀疑的东西。所以,欺骗看起来是要冒着减少被试合作的危险(Ortman and Herting, 1998),而且,就像是雷森和罗安(Resson and Rowan, 1981)提出的,"好的研究意味着永远不说对不起"。

压力和不适——被试的保护制度

心理学的调查应该确保被试的安全，做到保护他们远离任何可能的伤害或不适，对于这一点原则是没有争议的。困难在于试图确定什么样的压力或不适是生理或心理上无法接受的。人类学家等也许会声称在这类“主题”上的任何传统的实验性研究都是对人类尊严的侵犯。在一个不太极端的情况下，那些看到人类实验过程中价值的人也仍然会批评某些研究者太过分。

心理压力

在前面已经给出了涉及一定程度精神压力的研究案例。这些研究包括对人们的自我形象的丑化或是应对实验行为负责的过度疲劳的感觉，如拉托尼和达利（Latane and Karley, 1976）的研究。

并不是所有的精神压力都源自于欺骗。也许来自于向被试放映情色作品或是暴力电影的结果。被试以一种错觉和幻想的形式经历精神上的极度不适，遭受过“感觉剥夺”（剥夺听觉、触觉、视觉）的被试，他们通常会在三天后结束这个体验（Jolyon,1962）。津巴多（Zimbardo, 1972）的模拟权威和顺从者实验原本打算进行14 天，但在第 6 天就不得不结束。扮演犯人看守的学生表现得太逼真了，他们对待“犯人”变得极具攻击性、虐待性、野蛮性。他们的犯人（其他学生）变得极其消极和依赖。两天内或是在接下来的几天，被试不得不退出，因为他们开始表现出严重的情绪和精神混乱的迹象（无法控制的大哭和尖叫），有一个甚至是得了神经皮疹。

在此，调查者的责任就不仅仅是描述任务，还要尽量消除心理学研究过程中给被试造成的消极的甚至是长期的影响。在实验 1 年后，40 名米尔格拉姆实验中的被试接受了精神病医师的检查，精神医师报告没有一个被试因为他们的实验而受到心理上的伤害。1992 年的英国心理学会的原则中便强烈要求在那些被试可以接触到调查者的过程中，调查者必须通知被试在这些过程中可能会有压力或在参与后受到其他不好方面的影响。

身体的不适

许多心理学的实验运用了各种各样的手段，例如，电击，强度极大的噪音，食物和睡眠的剥夺，让人服用产生焦虑或是恶心的药物，等等。

众所周知的是，华生和雷纳（Watson and Rayner, 1920）就是通过在小艾伯特（Albert）跟小白鼠玩弄的任何时候都制造巨大的令人不安的噪音使“小艾伯特”这样一个小婴儿对一个他原来游戏得很开心的小白鼠表现出焦虑。小艾伯特甚至已经明显地开始对其他有毛皮的东西感到小心翼翼。在小艾伯特症状可能变得更严重之前，他们家搬走了，因此艾伯特也退出了这个课题。这个实验发展成为“令人反感的条件反射作用”可以使自愿的来访者消除他们不希望的或是消极的行为。

“自愿”这个术语会带来困难。在一些敏感的案例中，男同性恋使他们自己接受厌恶疗法（现在也在使用），这些治疗被认为是不道德的，因为这些男人屈从于传统正常的社会结构，认为他们自己的性取向是不可取的或是病态的。在通常的研究工作中，一个“自愿”的被试通常是在社会的压力下行动的。他们也许觉得他们破

坏了实验或是使试验者失望(调查者这种特殊的力量将会在下文讨论)。因为这些原因,调查者对于被试有一系列的责任来确保他们没有遭受过度的或是不必要的痛苦。这些在紧接着的部分会被标记出来。对于任何已预期到有令人不适的实验,在实验进行前,调查者都有责任向态度中立的专业同事寻求意见和建议。

同意和非被试的权利

在所有涉及个人参与的研究中,调查者必须做到以下几点:

1. 给被试全部的信息,包括可能的不适的水平,强调志愿者自然的表现和在任何时候撤出的权利。
2. 根据这条消息,从被试那获得他们经历实验过程的知情同意。尽管完全告诉信息不太可能,因为完全的信息会使他们的行为向着假设努力。他们可以获得足够的信息去作决定,例如"你的饮料可能含有高咖啡因"。
3. 提醒被试退出的权利,如果出现了比预期中还强烈的不适,在实验过程的任何时刻他们都有权利退出。
4. 当不适的水平远远高出预期或是被试明显被这种无法接受的水平扰乱的时候,可以结束过程。

现在我们应该知道米尔格拉姆的研究中许多惹人争议的方面中的一个了。实际上他测试的是实验者对于被试的影响力。实验者确实公然地违反所有这些原则。美国心理学会和英国心理学会都强调研究中尊重被试退出的权利和提醒被试所具有的退出的权利。与这个原则相反,在米尔格拉姆的研究中,而且是在一个也许不太人道的时代,为了研究计划的利益,命令每个希望停止的被试继续进行。当被试尝试拒绝反应,试验者也许使用"继续是'绝对必要'的"和"被试别无选择除了继续"之类的话。美国心理学会和英国心理学会也强调当调查者处于一种对于被试有巨大影响的位置时该有的特殊警觉心。当然,这也是在米尔格拉姆的研究中有了有力的发现并且得到了验证。

我们可能在下面发现,在研究操作前获取被试的知情同意并不一直是一件简单的事情。实验室实验必须得到被试的同意,尽管被试不一定完全知情。对于孩子的研究,必须首先得到家长的知情同意。原因很明显,孩子不太容易承受大的压力,即使父母会同意(尽管小艾伯特母亲同意)。对反对知情同意起作用的两个因素已经做过讨论,在某情况下可能是调查者的欺骗需要和调查者角色所附有的巨大的力量。

调查者特殊的权利

然而,通常在实验前和实验过程中,调查者必须给予每个被试不参与的机会。就像我们刚刚说过的,反对这个观点的主要是有影响力的调查者的权力。托伯特(Torbert,1981)说:

……只有在一种特殊的社会背景和专制的政治背景中,片面控制研究对象就是其本身。这不应该感到惊奇,一些最吸引人的周密发现就是关于人们对于独裁主义的反应。

当我们考虑到美国普通心理学大学生的时候,这种权力的一个额外的维度出现了。这些大学生往往有成为研究中被试的义务,尽管他们可以选择参与哪个研究。通常他们也可以不参与但是却要多参加一门期末考试,使这个选择流于表面。现在这个系统开始出现在一些英联邦的大学。

非自愿的参与——一些特殊性质的观察研究

在参与观察研究和自然的(隐蔽的)观察中,被观察的人对于他们的参与经常是无意识的。这看起来难以反对,因为在这些研究中,进行的是不打扰的观察而且每个被观察的人也只是一个计算的频数:例如,当观察司机时,目的是想确定在路上"停车"标志前是有更多男性还是女性司机停下来。一般的规则是如果这个行为是任何人都可以观察到的,而且因为这些内容又是公开的,那么记录下来也是没有问题的。

在参与观察的研究中,人们的私人空间也许会受到侵犯。汉弗莱斯(Humphreys,1970)通过在一个公共浴室"注意"的行为的观察,调查了经过同意的同性恋者的行为。被观察的人完全没有意识到研究的进行,而且通过后来的采访,事实上记录他们的汽车注册号码是为了获得更多的背景信息。

一些领域的研究是在公共的活动场所进行的,并且涉及一些干扰人们生活的操作。一个街道的调查明显地耽误了每个回答者,但是这些都会首先获得对方的同意。罗丁和皮利艾威(Rodin and Piliavin,1969)的对于旁观者干预的研究,一个人要么是跛脚,要么是喝醉"崩溃"在地铁上。在一个版本中,表演者咬了一个胶囊然后制造出下巴上慢慢淌血的状况。可以预期的是,这个"跛脚的"人会比醉酒的人得到更多的帮助。这种"流血"状态也会降低帮助的效果。事实上,皮利艾威等人的研究违反了公开的(没有欺骗),避免压力和在参与前知情同意的原则。"被试"无法决定是否参与。

杜布和格罗斯(Doob and Gross,1968)在一交通灯处使用了或漂亮的新车或老旧低价的车来阻挡司机。结果是可以被预期的:被漂亮的车阻挡的司机将要花费更多的时间鸣笛。如果这些结果令人感觉非常正常,难道不能要求自愿的被试简单地想象一下这个情景并考虑下他们可能的反应吗?在这种情况下,能否运用模拟情景呢?杜布和格罗斯也使用了问卷调查,并且发现独立的学生样本在对任何一辆车的鸣笛时间上的预期是没有差别的。而有一些人说他们不会对任何一种车进行鸣笛,在这些人中一个奇特现象出现了。当然,所有不会对老旧的车鸣笛的6个人都是男性,而所有不会对高级轿车鸣笛的5个都是女性。这个"好像"的发现跟事实上的行为是如此的不同,以至于研究领域的辩护者看起来更维护他们声称更加现实的数据。然而,到1991年,人们设计了一个电脑模拟情景而且这个程序证实了原来的发现(Bradley,1991)。

干 扰

上文提到的,那些因为轻信研究者,而受到干扰的被试的任务报告的一些方面都被处理了。虽然有自愿的被试,但一些研究中涉及较高程度的干扰。例如为了证明父母刺激对于孩子们的学习和认知方式的有益影响,心理学家会在家同父母和孩子一起工作(Klein,1991)。这样的研究中,经常用比较的控制组来作为基准是必须。在医院新药的实验中,实验也会被束缚,如果成功是显著的,那也将是不道德的,因为对于使用安慰剂的控制组的病人保留了他们使用新的治疗药物的权利。不幸的是,在心理干预的研究中,即使成功是明显的,通常这也没有一种政治上的力量和财力对所有贫困的家庭实现这种"治疗"。因此,在选择一组接受特殊治疗上,道德问题产生了。

在只有为了研究目的时候才能进行干预,并且当涉及通常被认为是社会无法接受的实验行为的时候,一定要慎重考虑道德原则。例如,莱恩等人(Leyen et al. 1975)的研究,那些观看暴力电影的男孩攻击性水平提高了。观察到这些男孩在日常生活中比那些观看非暴力电影的控制组的男孩更具有攻击性。怎么去单独汇报把这些男孩恢复到他们原来没有开始学习时的状态,这看起来是相当困难的。

13.4 心理学研究中动物的使用

动物爱护者将会很高兴听到,在心理学的纯理论研究中动物的使用将会大量减少。使用动物的一个论据就是它们可以被利用而人类不可以,尽管这样的说法相当于回避了整个问题。也有人认为在一些基本的行为方面人类和非人类都是相同的,在实验室的学习实验中,可以对各种情况实施更好的控制力,而且与动物行为的对比也许可以为人类行为研究引入新奇的视角(例如,就像是依恋理论的产生)。

反对使用动物的言论是完全道德的。例如,动物参与的实验研究是对我们应该努力尊重和保护自然宇宙的一次打击。这些反对言论也可以是从实用性角度出发,主要关注于从动物身上获得的知识缺乏价值,因为就某些方面而言,它们跟人类是如此的不同。例如,本能反应的数量、缺乏语言和他们对于特定行为的准备而不是其他。

在英国对于动物的研究操作要在英国心理学会发布的方针(2000)的指导下进行。在这里,介绍以下几点:

- ❑ 获得的知识必须证明过程的合法性;不鼓励实验性的研究;还有可供选择的方法。
- ❑ 应该使用尽可能少的动物数目。
- ❑ 濒危物种,不应该用来做实验。
- ❑ 对特殊物种的研究过程中,如有关进牢笼,食物剥夺,引起不适和痛苦的环节,都应该进行评估。
- ❑ 自然主义的研究更喜欢实验室研究,但是动物的实验应该在野外并尽可能少的干预。
- ❑ 实验者必须熟悉麻醉法、药理学化合物等技术性的方面;常规操作后必须做医疗检查。

13.5 结　论

总的说来,进行许多研究却根本碰不到争论性的道德问题,这看起来是很困难的。当然了,在进行任何实验研究之前先考虑到可能的道德方面的反对意见,这看起来也是不可能的。其他的自然科学也有学会和委员会考虑在科学研究上的社会责任。他们讨论这些发现可能实施到的地方或是研究从那些机构获得赞助这都是不谨慎的。他们要考虑他们的工作对于整个社会可能产生的影响。

心理学家必须有相似的考虑。然而,自从人类在社会中作为单独存在的个体,就已经成为了研究的聚焦点,作为一个研究公共圈,心理学必须非常警觉地发现不当治疗,虐待,欠考虑的和缺乏职业精神的情况,这没什么令人惊奇的。如果,心理学家不想让人们在一个聚会上后退,那么当他们说"我打赌你在考验我""真是实验的一部分吗?"这样的话时,心理学家需要不断地向公众保证一些过去过分的行为现在不会再次发生,并且真正的欺骗只有在必要的时候才会使用。

人类学者和定性研究者看起来在这些道德问题上已经获得了道德的制高点,不仅仅因为他们把尊严、诚实和人道主义放在首位,而且因为他们认为他们的参与或是非指导性的方法是获得真实不受胁迫信息的唯一途径。以里森和罗安(Reason and Rowan,1981)的研究为依据,马斯洛提出,"……如果你像对待东西那样去刺激人们,他们就不会让你了解他们"(Maslow,xviii)。

好了,你是怎么想的?

你也许正和你的同学或是同事非常热烈地讨论这些实验操作中正确的和错误的地方。我情不自禁地感觉到从米尔格拉姆的工作中获得的信息是极其有价值的。它当然也渐渐地否定了我的成见,而且很长一段时间整个文化倾向于更加顺从或是能够伤害别人。但是我也情不自禁地立即想到了这些尽他们最大努力的被试。难道我们可以保证我们就是那些停止的35%中的人吗?即使所有这些停止的人也是当受害者明显处于困境时才停止的。对于我们剩下的人生我们感到了什么?难道我们要使其他人承担这种尊严损害的后果?关于许多心理学家争论和哲学上两难困境的问题,我还没有得到任何最终的结论。幸运的是,我并不是必须要发表我的看法。但是你是怎么想的呢?

推荐阅读

Oliver(2003) *The Ssudents' Guidp to Research Echics*

练　习

在下列提出的虚拟的研究课题中,涉及的主要的道德问题是什么?

1. 研究者安排一个演员在街道上摔倒了,看起来很严重。研究者感兴趣的是在这种情景下路过的人是否比演员口吐鲜血的情景下停下的少?
2. 一个学生的第三期项目,她决定操作一个纵向研究的案件,研究中她会偷偷记录一类同事的言行和举止,这类同事是她认为正处于厌食症发展阶段的。

3. 研究者认为互联网的频繁使用跟抑郁的强度有关。她让学生去完成一个关于互联网使用的调查问卷，并且完成一个对临床抑郁的大规模的测量。
4. 让实验中的被试去完成一个涉及解决逻辑难题的任务。告诉组中的一半人他们做得非常好而告诉另外一半人他们做得相当的差。然后让被试评估他们在任务表现的内部或是外部的归因程度。
5. 机密性和匿名性之间的区别是什么？
6. "知情同意"是否意味着在参与前完全告诉被试？

答 案

1. 非自愿参与；知情同意；欺骗；精神上的压力。
2. 缺乏知情同意；非自愿的参与；缺乏在这个领域的专业知识；缺乏普遍的尊重。
3. 做什么的问题应该是被试在抑郁程度上的得分结果很高；研究者也许没有资格提议或是提供任何种类的专业支持；因此这是一个严肃的任务报告描述的问题。
4. 欺骗；心理压力；需要对"表现差"的组小心地任务描述。
5. 机密性意味着心理学家向来访者或是被试许诺，他们给予的任何信息不会被其他任何人看到。匿名性意味着这样的信息也许会发表，但是关于来访者或是被试的身份不会向任何人透漏。
6. 不是。研究是关于什么的，这个问题要给被试充足的信息，因此他们可以决定是否参与，但是为了防止被试的行为不受预期的影响，一些信息也许会被保留。

关键词

匿名(anonymity)
机密性(confidentiality)
任务报告(debriefing)
欺骗(deception)
知情同意(informed consent)
干预 (intervention)
非自愿参与(involuntary participation)
撤回的权利(right to withdraw)

14

实践研究的计划和研究报告的撰写

本章内容

本章主要介绍实践研究的设计、开展和撰写小篇幅的实践研究报告。

- 指导学生如何寻找一个合适可行的课题并形成一个可验证的假设。
- 针对你所做研究的具体的设计特点，我们将从以下几个方面给出建议：如何获得和处理被试，如何组织材料，控制实验进程，最重要的是道德方面的考虑。不要做任何与英国心理学会道德规范相违背的事情。
- 探讨研究报告的撰写问题。研究报告应是陈述实验过程的，因此通常用过去时态。所有的部分都应该有写作目的。文中所有列表或者章节都不应出现只有数据而无文字说明的情况。
- 对惩罚严厉但仍风气日盛的剽窃问题的有关建议。
- 详细介绍实验报告的各个部分。
- 最后，我们呈现一个中等学生的研究报告，这篇报告已经被评论并在上面做了标记。接下来介绍一篇研究同一领域的优秀的研究报告与前面中学生的研究报告相对照。

14.1　实践课题的计划

心理学课程中可能会包括一些对实践研究的评论。这里你将设计并实施一个实验调查，通常还要收集数据并提供描述或者推断统计。写作的时候，我们应该遵守所有A级委员会和国际文凭组织（the International Baccalaureate，IB）所规定的内容，但所有提纲对于写作特征和要求的规定都过于冗长而无法给我们具体的指导。因此，我们要讲一些常见的特征，提供一些建议和辅助措施以有助于人们更好地理解这些要求。

当开始一个研究或者开始提问被试时你的实践研究并没有真正开始，这只是所有的程序中很小的一部分。你应花费相当比例的时间去计划，并花很多的时间去分析（不知这么说是否妥当）和撰写实验报告。

关于实践调查

进行实验研究的一个好处就是：你有了一个可以真正进行心理学研究的机会。你可以对产生的某一想法进行检验及结果分析，并观察你所预期它在人们身上所能发挥的作用是否确实如你所料——当然所有这些必须要在伦理道德允许的范围之内。下面是一位学生为一项实验研究选择课题向老师汇报时所选用的两种不同方法：

A.“我想要研究关于心电感应（telepathy）的东西。”

B.“我对贝洛夫的研究中展示的人们通常认为父亲的智商要高于母亲的现象颇感兴趣。我想知道老年人群体和成长过程中的孩子是否都有这种看法。我还想知道是否哥哥们都认为他们的弟弟智商不如他们高。”

到底哪些学生具备这样的特征呢？他们有着明确目标并对自己想法的实施有所预期，从导师那里得到较少的意见，但可以从自己设计的细节开始实施自己的研究。哪些学生会有可能用相关并易于操作的假设来验证自己的想法呢？哪些学生的学识已经覆盖了教学大纲中所规定的心理学领域的大部分知识呢（如社会认知、刻板印象、性别）？

通常你会被要求用你所学理论的一部分来进行一个课题研究。因此如果导师对你选择的课题皱眉的话，你也不必感到不高兴。可以肯定地说，你必须要通过导师的审查才能实施自己的想法，而且你应该好好听取他们的意见，至少他们做这方面工作的时间要比你长得多！

从这点出发，本书的教学大纲中对于导师提出的关于实验目的的建议和一些设计的细节内容有所省略。如今，以英国资格评估与认证联合会（AQA）2007年开始执行的A2级的操作说明为例，由于导师提出研究假设和研究的设计而丢失的分数仅占60分中的3分，或者说只有5%。如果你认为这是很大一部分，请记住这仅仅是你课程单元内容的1/6。这部分仅仅是整个A等级分数的15%。因此，如果不设计你自己的实验让你失去的仅仅是15%中的5%的分数，也就是0.75分。你需要权衡的是总体的0.75分和如果你决定提出自己创新的想法并完全由自己设计实验

可能会失去的分数之间的关系。很少有学生可以看到所有易犯的错误,能够预期所有实验设计中可能出现的问题并能够发明一种能够精确地检验所提出的假设的测量方法。在只是丢失0.75分而非由于假设不充分和设计缺陷丢失更多分数的情况下,你最好对导师提出的指导意见全部接受。

多数教学大纲允许学生在进行研究时以小组学习的形式设计,但同时,你将会为此丢一些分数。然而,如果你在撰写学习报告时没有使用自己的语言,你将失去更多的分数。我们将在后面的内容中谈到剽窃问题。剽窃意味着有些研究成果表面上是你自己的工作结果,但事实上至少其中一部分工作是由其他人完成的。在你的研究报告中包括了一些其他同学的研究成果。想象一下这样的情景,你和其他人一起工作(原则上是个令人欣赏的方法——多数的科学成果是人们共同发现的),并且确信在工作结束后,你拥有了足够的信息来撰写自己的研究报告。

在你想要(或者不得不)寻找你自己的研究课题时

在听任何讲座时都带上一个便签本,当有趣的事(尽管有些事比较简单)发生时做个记录。或者通读你的课本并在其中查找一个简单而又有趣的研究。想想如何重复这个研究或者至少能重复这个研究的一部分,如果能做的更好的话,看能否改进这个研究,例如采取不同的研究材料,不同的被试人群(例如不使用学生被试,或者不在美国进行实验,或者实验时间不是在1956年)。或许你可以对原来的研究进行更加激进的改编,例如你可以用足球球迷做被试来验证归因理论(attribution theory)①中的自我服务(self-serving)偏见:分别在自己喜欢的球队赢球和输球后问球迷他们喜欢的球队为何会有比赛时的表现。根据归因理论,他们会将球队的胜利归为**内因**如球队的技术很好,而把比赛失败的原因归为外因如运气不好或者是黑哨(bad refereeing)等。进行的研究对你来说必须是切实可行的,例如不要进行关于日本和韩国依恋行为的跨文化研究(cross-cultural study)。

你也可以在公开发行的刊物上寻找灵感,如《心理学评论》(*Psychology Review*)、《心理学家》(*The Psychologist*)、《新科学家》(*New Scientist*)等。通过在网络上寻找或者在搜索引擎里键入"心理学杂志",你可以看到现在许多文章的摘要。以应用心理学期刊(*Journal of Applied Psychology*)为例,在内容框里你可以看到搜索结果所提供的所有文章的摘要和一些特选文章的全文。如果你的导师可以为你提供机会进入当地的大学的话,你也可以在心理学摘要数据库(*Psychinfo*)中查找资料。同时可以通过其他人感兴趣的媒体来激发自己的灵感,但是**必须**要记住:你的想法要能够追溯到**相关的**前人的研究或者理论。你的导师能告诉你什么样的研究同你的研究相关度比较大。不论在什么情况下,千万不要说你的想法是前所未有的(到目前为止,还没有人能证明这个预言),对此我们无从得知确切的信息,但几乎可以肯定地说这一想法曾在某个时候某个地点被某个人提及过。

形成研究假设

这里有个简单的黄金法则(golden rule)——不过在困境的时候可以不按此法

---

① 1982年Winner提出的归因理论,将事件发生的原因归为三个维度:内外因、稳定性和可控制性。——译者注

则执行：

在开始收集数据之前，必须知道研究的预期目的和如何准确分析研究结果。

认真地考虑你所预期的事情中哪一件会发生以及你将如何证明它。如果你还不确定怎样分析你的结果，请先不要开始这个研究，到导师那里去寻求建议。如果每次看到一个几乎发狂的学生拿着下载来的数据却不知道如何分析它们的时候，我想能得到一英镑的话，我想现在我应该很富有了吧。这样的情景可并不好受，希望你们可以完全避免这种处境。

我们已经知道，假设是对特定人群的一个声明。尽管课程大纲上对假设的提法也是从上述观点出发的，但是我们这里谈到的假设却是你对研究的预期。例如，你可以预期"男性可能认为他们的智商高于女性"。提出假设时必须使用具体的可操作性定义（specifically-measurable terms），而且这些操作在你收集数据的过程中必须具有可行性。例如，也许你需要对"关怀""同情"和"自尊"的不同之处作出一个预期。这些变量该如何测量？你可能会在书上或者因特网上寻找到答案，（如一般健康问卷（Goldberg，1978）；或者罗特（Rotter，1966）的心理控制源量表）但是作为实验设计的一部分你需要有自己的创造发明。

有时，你可能会想要改变自己的研究目的或者研究假设的侧重点，这要取决于何种测量方法是可用的。你需要确保在你的引言部分里有足够的研究文献来支持你的假设。如果你的假设或者预期是在"常识"的基础上得出的，这将不会为你的报告赢得更多的分数。

调查的设计

我的许多学生在最初计划他们的实践研究时都以"我想用一个问卷"开头，对于这种问题我通常都用"为什么"来开始同他们的交流，结果是他们大多数都被问的一脸茫然。问题在于，只有等你意识到自己真正想要寻找的是什么，你的整体设计是怎么样的和你将采用和/或发展什么样的测量方法以后，你才能确定你是否需要一个量表。

你的设计就是你在构架你的研究以证明你的研究预期时的整体思路。例如，如果你认为焦虑会同自尊呈负相关，这时就要求你使用相关法（correlational approach）进行研究并找到测量焦虑和自尊这两个变量的方法。此时你要寻找一种实验方法，并确定这是一个真实验研究。例如，在标准皮亚杰日常守恒（conservation）的研究中所显示的幼年儿童倾向于认为定量的橘子汁在不同的容器里时其体积发生了变化，就不能称之为一个实验，这个研究中没有自变量（independent variable），只能称做一个验证罢了。如果你要比较年幼和年长的儿童之间的差别你可以引进一个自变量，但此时它仍不是一个实验变量而是一个群组差异研究（见第4章）。

不要引入太多的自变量

你需要处理的变量是否过多了呢？假如你要看性格内向的人在独自相处和在公众面前时相比对同一个任务的完成情况是否有所提高，性格外向的人在两种情况下相比任务完成情况是否有所降低；你可能想要知道这种情况对男性更有效还是对

女性更有效，你可能想要知道年龄是否也是一个影响因素。考虑到变量的交互作用，这样将使统计分析变得非常困难。这个级别的教学大纲将会指导你采用适合仅有两个水平的单变量的推断统计（除非你打算使用卡方检验）。

重复测量

你的实验能使用重复测量法（repeated measure）吗？如果可以的话，那当然很好，因为整个实验你只需要很少的被试并且不用担心你的被试样本会异质然而，有些时候需要采用单盲实验的设计。如果你采用的是情景故事研究（vignette study）（见专栏 14.1），那么采用重复测量的方法就没有任何意义，因为情景会使被试产生身临其境的感觉。你需要使用独立样本设计（或者“独立的测量方法”）以达到使每个被试只经历一种情景的目的。

专栏 14.1　采用情景故事法的实验设计

我们经常会使用一条信息来作为我们的自变量。例如，我们可能想要知道人们对引起严重后果的肇事司机是否责备得更加严厉。很显然，我们不可能对这样的研究进行真实验设计，而且为此制作一个短片也是既困难又昂贵的。这里我们能用的就是情景故事研究——让被试阅读一个对特定场景的描述，然后让他们作出一些判断——例如，对人们认为汽车司机对事件应付责任的程度进行数字化评定。

沃尔斯特（Walster，1966）的研究证明被试的确会认为造成破坏越严重的司机所负的责任也应越大。随后，麦基利普和鲍萨维克（Mckilip and Posavac，1975）的研究也证明被试和事故受害者的相似程度也会影响被试对事故的严重性和事故责任的判断。同没有使用过大麻的被试相比，有吸毒经验的被试认为使用大麻的肇事司机（虚拟的情景）即使在严重的交通事故中应负的责任也较少，判定较少的罚金并将事故归结为外部原因。实验支持了认识上的相似性会降低对责任的归属这一理论。

很显然，你不能询问被试对药物的使用情况但你却可以通过把肇事司机分为成员组和非成员组（non-number group）如，一个学生，一位女性，一个慢跑者等来验证与其相类似的假设。

通过使用情景故事法你可以验证：

1. 一篇文章将其作者标为男性时和同一篇文章将其作者标为女性时人们对它的判断是否会受影响。（Goldberg，1968）。
2. 当阅读一篇关于一间房子的故事时，和普通的买家相比，夜贼对屋内贵重物品的记忆（如油画）要多于对普通物品（破漏的屋顶）的记忆。（Pichert and Anderson，1977）
3. 一个**太阳报**的虚拟的读者对一起交通事故死亡描述的句子是否长于一个**卫报**读者的描述或（者相反）。[1]

……

---

① 《太阳报》（*Sun*）是全英国销量最高的报纸，虽然它的销量、知名度非常高，但不少人对《太阳报》都持有不良好的印象。批评《太阳报》报道新闻的手法粗糙、不专业、不中立，常常以哗众取宠的手法来刺激销量，读者对象基本都是学历知识有限的人。《卫报》（*Guardian*）是全国性综合内容日报，其风格为：严肃、全面、客观地报道新闻，对重大事件阐发自己独到、客观的观点。——译者注

自变量的隐藏

在一个情景故事研究或相似的研究中，如果想确定被试是否已经注意到作者是女性，或者司机是学生等，便可以开始提一些阅读前的预习问题，如作者多大年纪了，他们住在哪里（在刺激材料中的确已经出现过这些信息），然后逐渐过渡到关键性的问题如“作者的性别是男性还是女性”。这样可以避免引起被试对此项实验变量的怀疑。

考虑变量

尽管我们的测量方法都是在“给定材料”的前提下讨论的，但是对变量测量水平进行认真的考虑还是明智的做法。如果你需要进行相关研究，最好不要选择分类变量。如果需要进行一个性别和焦虑的相关研究那就需要再三考虑了。你必须理智地做的事情就是，检验不同性别样本在焦虑水平上的差异。因此，如果要设置不同的变量水平就需要一个合适的量表。这样就不是要记录司机是否停车而或许应该记录他们的速度了（在路上两点间所用的时间）。

观察法研究

如果你要进行一个观察的研究，需要非常的仔细，因为当课程负责人将研究命名为“动物园一天的研究”或者“我可爱的宝贝表弟”时研究并未结束。你需要制作合适的观察表或编码系统（仔细计划自己观察的重点所在），安排一个观察操作员，因为要把有用信息聚集起来并将这些写入研究报告中是非常困难的。你的报告不能写得类似于一个记者或者一个动物爱好者所写的东西，你写作的内容跟他们越相似就离十足的心理学实践报告越远。

访 谈

如果你打算使用访谈法但是却想获得定量的数据，那么你就需要考虑怎样将定性的内容转化，除非你只打算使用一个特定用途的问卷（fixed-choice questionnaire）（如，只有封闭问题的量表），否则就需要考虑对定性材料的内容进行编码。如果想把访谈的内容划分等级则需要考虑请人进行独立的分级以避免主试的个人偏见对结果产生影响。为此你需要创立一套严格的编码系统，而且进行编码的合作者也要经过专门的训练。

访谈法需要大量的时间而且有可能需要仪器，因此要确定自己有能力可以采用这种方法！而且必须考虑时间上的可行性。

被试的获得与处理

- ❑ 你是否能够获得足够的被试实现自己的研究设计？被试之间是否能够进行匹配？你也可能难以获得你所需要的其他相关信息（如受教育程度）。
- ❑ 如果采用反复测量法在两种不同场合使用，那么第二次测量时，是否所有被试的数据都是可用的呢？
- ❑ 应避免使用难于找到的被试（左撇子，家中的第五个孩子等）

- 如果想开展群体差异(group difference)的研究(如不同性别的研究),你需要保证被试的相关变量在操作的层面上要尽可能地匹配。
- 如果需要控制组的话,就要保证控制组的被试除了在自变量的关键水平上跟实验组接受不同处理外,其他的操作都要和实验组相同。例如为了研究回忆类任务中被试表现的差异,在重听单词或在词表匹配图画判断时需要给被试足够的时间并尽量增加实验的趣味性,而且要给出一些事例让被试知道你希望他们做什么。在控制组被试完成实验任务时,我们也要花费同样的时间和精力以避免因投入不足而对实验结果造成影响。
- 你的被试是否是熟悉的朋友、熟人或者是餐厅里的学生?如果是,他们是否会将你的实验任务交给其他人代替他来完成?
- 你不能从人群中随即挑选样本(千万别说你可以做到)但是你可以对特定的变量如年龄、性别、资格等寻找到匹配的被试组。你可以随机把你的被试分配到不同的情况下。
- 如果你怀疑被试会搞砸实验或者已经知道了实验的目的并试图表现得更好,那么无论他们的反应如何合理,都应把他们的数据放后或剔除。遇到这种情况需要同你的同事或者你的导师讨论。
- 确保有足够的被试样本,我们大致的原则是:实验时,如果采取重复测量的方法样本量至少应该为10,如果是用组间测量样本量至少应为20,如果进行某种形式的调查,样本量则至少应为30~40。有些学生曾经很庄重地声称:他们用一两个关于种族容忍主义笑话的问题的研究结果支持他们的实践假设,即北方人的种族主义倾向要少于南方人(在英国)。这项调查的问题在于他们调查的对象只有8名北方人和5名南方人!

材 料

- 在两种情景下材料是否一致?如果当抽象类的单词长度整体上长于具体词时,那么比较具体的词是否比抽象词更容易引起记忆唤起的测试则是毫无意义的。我们应努力确保材料之间的差异不能同自变量引起的差异相混淆。
- 两个记忆词表是否能够相等呢?你是否能说词表里的所有单词在正常情况下使用频率相同或者一组人工词(anagram)可以产生相同的识记难度?对于这个问题你可以进行一个预测来证明材料之间并没有实质性的差异,或者你可以列出词频表,下面的网络资源将对你解决这类问题有所帮助:http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html
- 给被试的指导语是否易于理解?如果你对问卷中的措辞尚有所怀疑,那么请一个擅长使用语言的人帮助你吧。如果你的问卷中存在诸如拼写错误这样的问题,被试将不会慎重或者认真填写这份问卷。
- 如果你想自编问卷(参见第6章),请记住关于态度的测试通常不使用提问题的方式,而是用人们可能同意或反对,或者跟他们的观点相去甚远的**陈述句**。不要使用这样的句子“你是否相信堕胎/原子能/罢工?”这些事情早已存在,我们想知道的是人们对这些事情的看法。应该设计一些有共同回答形式的问题,并

给出分值,让每个人可以根据自己的看法给量表整体评分而非单独分析某个题目。

- 网络为心理量表提供丰富的资源,但是要注意的是许多量表可能出自外行或者业余爱好者之手,同时不要遗漏量表的评分系统。如果你非常渴望使用网络量表,可以参考美国心理学会的网站:http://www.apa.org/science/findingtests.pdf
- 如果你关注的是一些特殊群体的人们,如少数民族或者残疾人,请仔细检查并邀请研究该群体的相关人员或者其他的专家来斟酌你选用的语言。如用了不受欢迎的、贬义的或者居高临下的词组或语句,这样很容易让人觉得你不仅是无礼的,而且是无知的。这就要求我们无论何地,也无论该特殊群体的成员是否被提问,都应将注意力放在他们身上。
- 最重要的是要有礼貌!可以先在你的朋友或亲属那里试用你的实验材料。

过　程

- 如果你采集的数据中包括一部分学生,请不要由于很多的突然变化而使实验匆忙结束,让同事进行最后检查的行为会让你感到羞愧。如果你对你所做的事并不是十分肯定,不要认为自己很傻,酌情向同事或者导师询问意见。你肯定并不想由于你对原本定好的实验过程的误解使得实验数据最终难以汇总而使整个实验组终无所得。同时应该注意的是,当开始收集实验数据的时候不要忘记你的同事。如果你没有意识到这点,整个项目可能会被暂时搁置,使每个人的计划都不能按时完成,或者更严重的——最终使报告也不能如期完成。
- 记录现场所有的信息。如果你决定一直等会再去记录访谈对象的年龄或职业的话,你将会忘记一些内容。这样的话,结果可能就浪费了。如果看到相关的情况,记下这一被试是在哪个组,或者他们的性别。如果你忘记了,那么你的努力就白费了。
- 提前做准备使被试能保持轻松,并给出鼓励性介绍,为被试制定明确的指导语。与同事对实验过程进行模拟,你对哪些现象难以解释?人们可能还需要知道什么?
- 要想好如何回答被试可能会提出的问题。你是要对被试的问题早有准备呢,还是告诉被试等到实验结束后才告诉他答案?
- 如果研究采用观察法:

  ——观察是否唐突?提前检查记录的位置。

  ——记录是否困难?如谈话时使用录音机记录是否会过于吸引被试的注意力?编码系统是否有用?是否有足够的时间和空间做书面说明?

  ——为了获得高信度,我们是否可以同时对多个人进行记录呢?

道德规范

作为一个学生,你可能没有受过足够的训练而不能进行令人满意的述职会议。许多心理学专家经常讨论的一件事便是使人们回到最初的起点并"撤销"对人们的心理所造成的任何伤害。因此,你所提出的研究课题中**不能**包含以下几点,这是非常重要的:

- ❑ 侵犯隐私。
- ❑ 有损被试的尊严。
- ❑ 减少被试对自己的思考。
- ❑ 能引起怨恨和敌意的欺骗。
- ❑ 不必要的资料扣留。
- ❑ 痛苦或不适。
- ❑ 打破当地的禁忌(如饮酒)。
- ❑ 所有引起被试不适感觉的事情。

向被试保证匿名时,一定要保证做到真正的匿名! 即使是匿名的,如果对参加同一研究项目的同事或者好友以贬损的态度谈及被试也是极其无礼和不好的做法。如果你有上述行为,那么在那些试图帮助你的人中间你已经发展成为"精英人物"或者是"操纵者了"。[①]

同时,向被试保证不会让他们感觉到或者看起来很蠢,或者揭示任何他们不想让人知道的事情并向他们证明说到做到。向他们保证他们可以打破任何已有的行为记录,尤其是那些让他们感觉不好的行为。要提醒他们,如果他们愿意他们可以停止任何行为。

14.2 撰写实践报告

现在我们进行到了每位学生都喜欢的部分了——实践报告的撰写。我的第一个忠告是:**不要半途而废!** 当你最初的热情消失殆尽,当你不明白为什么要采取某项特殊的防范措施,或者不明白某种特定的情景是什么时,要坚持下去显得尤其困难。你可能会发现数据和分析时的基本细节材料找不到了,而你需要他帮忙的那位同班的同学也正因为原始数据的丢失而忙得焦头烂额根本无暇帮你。

正如你花费了几个小时完成了工作后,想想你所获得的那些能在你工作生涯里帮助你的那些一般性的技能,这或许会令你感到安慰。学生通常认为在他们大学的学位课程中学到的最有用的事便是学会了如何撰写研究报告。撰写报告这样的事可能以不同的形式出现,但是你总是要去做的。例如在你的职业生涯中,或者在进一步的深造学习中,甚至作为一地方社区的成员或者一个利益团体的成员撰写报告都是用得着的。报告的功能是可以清晰地同别人进行交流和沟通。

报告的目的是什么?——是关于你的经历的描述

写报告有两个目的但绝非是为了取悦你的导师。目的一是要告诉你的读者你做了什么,你为什么这样做,你认为你为知识的存储和理论的发展做出了什么样的贡献。目的二是将你的实验过程的细节都呈现给读者,使后人可以重复你的研究。我们也可以在别处看到为什么在科学方法领域研究报告会如此重要的原因。下面是撰写研究报告的第一黄金原则:

---

① 这里用的是反语。——译者注

> 确保你所写的东西有足够的深度，并且可以使一个对此领域完全陌生的智力一般的正常人，可以按照你的步骤重复你实验的每个细节。

实践报告是一则故事。这是你给其他人的一个门户，告诉别人你所想的和你要证明的东西是什么，为什么要证明它，你做了什么，你是如何做的，证明的过程中发生了什么事以及你最后的结论是什么。很多事情是依次而行的，其中最重要的是：

- ❑ 要使用过去时态。
  不要出现这样的句子"这个实验试图要证明的是……""被试将接受……测试"等。
- ❑ 确保读者知道你在说什么。他们当时并不在实验现场，他们并不知道你做了什么，而且他们也没有心理感应术。

例如，早期当你谈论"问卷"或者"情景故事"的时候，做个简短的停顿并自问一下"从读者的角度来看，我是否知道这是什么东西，我之前是否已经对此做过介绍了？"从评分者的角度来说，最让人沮丧的事情之一便是学生并不同读者交谈而只是纸上谈兵，因为学生们不得不这样。

我如何知道应该做什么

目前对于如何发表学术报告已经有了普遍接受的惯例，但是在细节上还有许多的不同。A 级委员会对报告的细节作了明文规定，你所参加的任何其他课程的组织者也应该这么做（如果没有，你可以要求他们）。大多数的规定可以帮助你组织你的陈述，使其更加的清晰和公正。从网络上找一篇文章（务必选自优秀的期刊杂志）并且/或者看你的导师是否同意你阅读以前的有着优良成绩的学生作品。在本章的最后我附上一些中等水平的报告及其得分评论。看一下你们考试委员会提供的资源材料，除了有点令人畏惧的书生气之外，史密斯（Smyth，2004）所著的《心理学写作原理》（The Principles of Writing in Psychology）将是一本绝对有用的书，它将在心理学写作各方面都对你有所帮助。

剽 窃

剽窃是指你将其他人的工作成果据为己有。包括占有他人的数据，其形式不仅仅包括完全拷贝，也包括你对别人的工作成果进行的近义转述。如作者的表述为"实验证明，在高温情况下被试所回忆的单词量明显多于低温情况"而你的叙述为"实验说明，在高温条件下的被试的单词记忆量要明显多于低温条件下的被试的记忆量。"你的这种行为便属于剽窃。剽窃将受到严厉的惩罚，由于剽窃本身是对科技的盗窃，因此我们应当如此对待。如果一个学生在考试时作弊，其他的学生会因此而得不到学位或者拿不到 A 等成绩，剽窃其实类似于考试作弊。重点在于，被视为个人成就的内容必须是自己本人亲自取得的。不要试图用下述的方法为自己开脱"我是从一篇文章的注释里抄来的，但当时我并未意识到这一点"。剽窃就是剽窃，不论你是否是有意而为之。无论怎样说从文章里直接抄袭都是没有任何意义的。从教育方面来讲，我们从抄袭中学习来的东西非常之少，而大部分的东西都是从心理学学习内容的记忆和学习过程中学到的。从伦理方面来讲，抄袭无异于偷

盗。这不单是指从书上抄袭,从网络作者那里抄袭也是毫无意义的愚蠢行为,这些作者对特殊的任务既不知道问题是什么,也没有一套具体的指导方针。当然你也不能凭空产生什么想法。所谓学习就是认真了解过去所发生的,然后希望自己能对前人的研究有所增益。然而,课程的重点是让你知道,你可以用你自己的话来表述要求掌握的概念。最好的过程是读书,自己做笔记,合上书,给自己提问并检验一下你理解了多少,然后试着将你现在对他们的看法写出来。如果你在报告中使用双引号你就需要给出所引资料的出处,这里要求具体到页码。这一点在实践报告的引言和讨论部分中的重要性不亚于任何其他类型的文章。当然在上述的两种情况下你都不应过分地依赖引用,你不应使用他们来避免对一些复杂事物的界定或者要花招试图通过用你自己的话对概念或观点复述的方式来表明你对它已充分理解。

一篇标准报告的组成部分

专栏 14.2 中列出了研究报告的标准标题。这些并不是一成不变的。你可以在杂志中任意挑选一篇文章,你会发现它的结构可能异于我们所给的模式。文章的结构因不同的杂志及其历史渊源而异,如"硬科学"(认知神经心理学)或其他种类的科学(社会心理学、健康心理学)。然而,如果要为本课程写一篇常规的定量分析报告,那么在你所能提供任何具体说明的场合下都可以使用专栏 14.2 所列格式。

专栏 14.2　研究报告的主要组成部分

报告中的标题	常规报告中每个标题下的内容
题目	告诉读者研究的领域及重点
摘要	概要
引言	主题领域的一般背景
	相关研究
	论点/ 论据
	总体目标
	研究预测("理论假设")
研究方法	你具体做了什么
实验设计	实验所用的研究类型、自变量、因变量及其水平
被试	参加实验的被试人员的具体信息
材料/仪器	设备等
过程	事实上发生了什么事,被试的指导语
结果	你的发现
描述统计	平均数、标准差等。包括图或者表格。
推理分析	用精确的结果和显著性水平进行推理性验证和证明。总结。
讨论	讨论结果,将引言中的研究放入具体的情景中,对研究进行评价,得出结论。
参考文献	在文中出现过的所有的作品及其细节都应出现在这里。
附录	对读者来说并非很重要的附加材料。

标　题

题目应尽可能简单明了。“一个关于是否能……的调查”这样的题目是不需要的。题目中只要体现主要的变量就可以了。通常在实验类的报告中可以用自变量和因变量做标题。例如，“图像法在语言材料记忆唤醒中的作用”此题目便可以将一个研究说明白（或许你看了也会觉得熟悉）。又如用相关法进行的一个现场实验用“在环境问题上年龄和态度的关系”这样的题目就足够了。避免使用问题或者喜剧性的题目如“男孩会高估他们的智商么”或者“观察侦探”。如果你在标题的前半部分确实无法抗拒使用这样的描述，那么在你标题的后半部分确保信息要充分翔实，例如“……智商自我估计方面的性别差异”或者“刻板印象假设在警匪电视连续剧中内容分析法的……”

摘　要

摘要应该用不同的字体和/或者缩进的方法区别于报告的其他部分。摘要最多在200字左右，这部分要体现全文的主要特点和整体框架。在这一部分不需要体现任何关于实验方法的细节，只需要讲明实验的理论依据，主要的设计思路，效果的显著性和突然发现的临界点或者结论等。你可以参考一篇“好的实验报告”并将它作为自己写作的样例：

究竟为什么我们开篇就要做一个总结呢？相信你会对是否有人在你提出的课题上有所行动感兴趣的，你的课题为：东伦敦红胡须的素食主义者的焦虑和慢跑行为。当你浏览了众多相关作品后，你会发现如果在文章的开头而非结尾就看到了作者得出的结论将给我们带来多大便利。

引　言

我通常将它形容为一个漏斗：

以心理学课题领域为开始。讨论和研究课题特别是与你的理性推断相关
的心理学理论和研究作品。从研究发现转移到你的
研究操作的辩论上。先做一个整体介绍，
用前人的研究结果讲你的研究假设
通过一个明确设计
把你要检验的
假设联系起来。

你的研究预测必须是从一个基于前人研究的结果论点或者在你引言的前一部分已经介绍过的已有理论（我们通常称为“理论概述（rationale）”[①]）中延伸出来的。如果在你所叙述的前期研究、基本原理和你的研究预测之间没有明显联系的话，你将会为此失掉很多的分值。你的资料可以来自于课本、研究论文、电子数据库和网络等，请注意对所有你将要引用的资料都标明出处。如果在读后很长一段时间后才

① 这就是研究性论文开始部分的理论阐述，说明随后研究的目的和原因——译者注。

去追溯他们的出处有时是非常痛苦的。

作为漏斗理论要求的一个例子,我们来看一下用图像法学过的词表里单词唤醒情况的假设验证的实验报告的引言。要进行这样一个验证假设,引言里肯定不会包括一个5页的记忆心理学的文章,也不可能包括艾宾浩斯的研究成果和法庭上一位证人的证词。假设验证属于一个特殊记忆研究领域,争论的焦点是复述(rehearsal)是否是短时存储记忆向长时存储记忆转换的核心的和必要的机制。我们可以通过下列步骤将读者从引言引导到正文:

1. 短时记忆储存和长时记忆储存的概念。
2. 两种记忆模型过程的提纲。
3. 模型可以解释的现象,如在自由回忆任务中的首因效应(primacy)和近因效应(recency)。
4. 重点在于模型强调把复述作为短时记忆存储向长时记忆存储转化的中间过程。
5. 介绍下人类总是试图在感官数据之外建构"认知"的缺陷;举例说明。
6. 从这个理论出发,尝试通过视觉呈现词目,并进行连接来给一个不相关的词表赋予某种意义的方法,在记忆实验中的效果要优于我们假定让被试通过阅读学习而进行复述所产生的效果。
7. 另外需要注意的是,要指出前期相似的研究和研究成果的学术水平。

我们已经讨论过,并为提出明确的实验预测作好了准备。需要注意的是始于本书第228页的"平均水平"的报告中引言相对松散,而在相关的文学作品中则是比较严谨。

需要避免的是我所谓的"**悬崖边缘坠落**(falling-off-a-cliff-edge)"综合征。它指的是作者在一般性的课题上谈论了很多,突然并无端地出现了关于记忆研究要验证的假设,并且在报告里既没给出解释也没出现相关的基本原理。这种情况会使你大量失分。

假 设

这里我们要用最清楚的语言来陈述我们的实验预测(假设),使读者可以清楚地明白我们期望的结果是什么。尽管传统的量化研究的结果的确会验证提出的假设,但是在报告引言的结束部分通常还是要给出一个明确的研究预测。例如,在一个图像实验里,假设通常是图像可以提高被试的记忆唤醒水平。然而,为了说清楚假设是如何被证实的,我们进行实验设计(至少会列出实验提纲)并作出预测:如果产生了什么样的结果就说明我们的假设被证实了。或许在某些课堂上你的老师会要求你写出实验假设,但是通常研究论文中引言的最后会将研究假设作为该部分所应具有的内容写出来。但是这里没有一个单独的"实验假设"这样的标题。例如,在我们关于记忆练习的报告里,最后一段就可以这样写:

> *对词表里的词的复述不一定必须要进行语义的转换,然而,口语材料在使用图像时则需要进行深层次的语义加工。所以说如果更深层次的语义加工会使以后的记忆唤醒更容易的话,我们便可以假设使用图片组的被试将比仅仅使用单词组的被试记忆唤醒水平更高。*

如果你想把实验假设单独写出来，那么上面段落里的黑体字就起到这样的作用。虽然我们通常将这样的句子叫做研究预测，但它实质上就是报告的研究假设。

或许你会遇到让你写出一个虚无假设的情况。这个奇怪的要求可能会让你产生误解认为这是为了要检验你上课的质量。在这一部分里没有人会写或者仅仅提及实验的虚无假设。正如我在第 9 章里所说虚无假设只具有统计学意义，并且在假设你的猜想正确时作为一个比较基线来用。统计检验的计算中已经包含了虚无假设的情况。如果有人要求你给出一个虚无假设，可以参考下面的例子或者对它进行修改就可以了：

> 在呈现图片和呈现单词两种情况下，如果在单词记忆水平上得分的平均数没有差异则接受虚无假设。

这个陈述是正确的但这还不是虚无假设，它只是你所期望的真实的结果。这个句子所表达的意思是“如果……，那么虚无假设成立”然而，实验者可以好好组织措辞来充分表达自己的意思，这样可能给他们赢得满分。（尽管你可以请你的导师写信给管理委员会改变课程提纲。）

这部分的关键是你的实验预测要用操作性的术语写得完全清晰明了。需要记住的是你最后的推论性结果分析要同你的实验预测相对应。“人们在使用咖啡因后记忆力会有所增强”这样的叙述是信息不清楚的。你必须对记忆力是如何增强的进行**定义**，如记忆正确的单词量有所增加。这样的假设或预测仍旧没有包括实验的基本原理。我们通常不采用这样的说法：“自尊和学术成就之间相关，因为人们成功的时候自我感觉会更好一些”因为这只是假设讨论的一部分。我们只能简单地预测，自尊方面的得分和精确的成就测量结果（如 GCSE 的数量和得 A 级的课程数量）之间的相关。

表 14.1 中给出了宽松和严格的研究预测的写法，左边是题号，你可以试着只看第一栏的内容写出其余两栏的内容。

表 14.1　撰写严谨的研究预测

练习题号	措辞散漫的研究预测	严格的研究预测
1	图片更容易唤起对单词的记忆	图片情景下记忆单词量的平均数要高于复述情况下单词记忆量的平均数
2	性格外向者在自信心上的得分将会高些	被试在外向性量表上的得分与在自信量表上的得分成正相关
5	在有观众的情况下人们在排序任务中将表现得更糟糕	在有观众的情况下排序任务用时的平均数要长于单独工作的情况
6	价格越贵的车在遇到交通灯时停下来的可能性要更大	在便宜、中等系列和高档车的列表中，从停到不停的比率将会增加

“比……多”的不可分离性

通常人们认为，当用“多于”时读者会知道你指的是什么，但事实上读者对此一无所知。在实验研究中请记住表 14.1 第二点建议。现在我们来看一个实验预测的实例：

> 我们的假设是：在盗窃情景中被试可以记住更多价值不菲的物品。

作者到底是什么意思呢？实事到底是怎样的呢？难道是我过于吹毛求疵了？

事实上，预测应为盗贼被试所记忆的价值不菲的物品数目是多于日常用品的数量呢？还是作者其实想表达盗贼被试和普通买家被试相比，对昂贵物品记忆的数量比较多这样的意思呢？这样模糊的表述导致我们实在不能确定作者想要表达的意思。请记住，当你使用“多于”或者“少于”这类词时，虽然有时你认为句子的意思是显而易见的，但还是要把比较的对象说出来并把整个句子写完整。通常在没有“比……”时是不能使用“多于”的。友情提示，如果你所指的东西不可数时用“less”，如果可数时用“fewer”，如更少的汤，更少的薯片等。

方　法

设　计

这里只简单地讲一下研究的基础结构——框架和提纲。如，所做的研究是否是一个实验，如果是，那么使用的是哪种实验设计（组内或是组间）？实验条件是怎么样的？有几组被试？使用每组被试的目的是什么（控制或者安慰剂（placebo））？每组有几个被试（这样的信息会在被试部分里提到）？多数情况下，对被试的描述也是对自变量的描述。不论在什么情况下，自变量、因变量及其水平都应该在这一部分作出明确的规定。另外，对于实验的进一步控制，如被试的匹配等虽然在实验过程中也会出现，但在这里也要进行说明。在我们关于图像的实验里，实验设计可以这样写：

> 实验采用的是重复测量法，每个实验组有15名被试，呈现的材料是20个项目的字表。情景一中被试的任务是借助图片来记忆单词，另一种情境中被试的任务是重复每个单词。为了对实验进行控制，一半的被试采用上述顺序，另一半被试采用相反的顺序。因变量是在自由回忆情况下被试记忆正确的单词数量。

这样就已经可以了。这里不需要给出实验材料或者实验过程的任何细节，而且对于后面会有详细介绍的内容不必在这里重复叙述。

如果是非实验性的研究，这里将要介绍设计的整体思路（例如观察的设计）和设计的框架，如纵向研究、横向研究等。在因变量和自变量中可能还有一些不可控制的变量，例如在十字路口等待的行人的数量和司机是否会停车等。观察者测量信度时，对测量可靠性的控制也可能会对实验结果造成影响。在这里只需要说明进行的实验控制是什么即可，不需要有细节性的介绍。

被　试

给出被试数目，包括每组中有多少被试，以及其他的和研究有关的被试信息。如果有人想要重复你关于青少年及其自我概念（self-concept）的研究，对于被试情况的介绍就显得非常重要了，比如被试的年龄、性别及来源。在技术性任务的实验中，这些变量可能不太重要，但是大致的年龄范围和左撇子或右撇子等相关信息还是很重要的。对某些研究来说，像社会阶层或者职业等变量信息将会非常重要。当然被

试对心理学的了解程度也是相当重要的信息。如何获得被试以及如何分组等细节应尽可能的详细。

除非你的被试是随便挑选的，否则不要轻易说你采用了“随机取样”的方法。“随机抽样”这样的说法是毫无意义的，它对于选择被试所提供的信息无异于说明这些被试刚好是方便选用的。对如何挑选被试的信息应适当地报告，即使当时的情况是你看到谁在现场就选用了谁（甚至其中还包括了你的男朋友或女朋友），也要坚持这样做。不要采用“被试来自于一个机会样本”的描述方式，要这样描述：“被试是一个机会样本”。

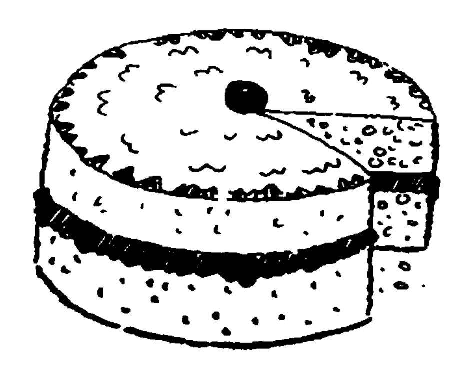


图 14.1　材料和过程部分正如一份蛋糕食谱

材料/仪器

在各种各样的教师工作坊里，我观察到卡拉·弗拉纳根（Cara Flanagan）将研究报告中方法部分的撰写比作写蛋糕的配方，尤其是其中的材料和处理过程，可以把他们看做是制作蛋糕的配料和说明。写作这部分内容的黄金法则是：给出足够的细节信息使得后人可以重复你的研究。这意味着要对商品类（迷宫（maze）、电脑）、设备（手指迷宫（finger-maze）①，错觉箱（illusion box）②）构造和设备资源（制造商、材料、型号）进行说明。重要的细节可以在这里给，也可以在附录里提供如词表、问卷、图画、情景故事等。但是对于白纸和铅笔这样的实验材料不需要给出细节。

不要简单地将所需材料列成一张清单。材料部分的写作也应像一般散文的写法一样。这部分同样可能包含一些有用的图片、表格等，或者在需要时对次序作必要的调整。但此处没有艺术造诣的额外计分。

在记忆研究中，我们需要两张词表，因为在缺少混淆变量的情况下，我们不可能让被试两次使用相同的词表。这里我们可以陈述一下如何匹配两个词表，如使用词频表进行匹配。所选词可以放在后面的附录里，但是必须要向读者说明如何可以找到这些附录。考虑到匹配词的来源问题，如使用网络上已有的词表，你就需要附上关于这些资源的所有参考资料。

参考资料里也必须包括你使用的任何公开发行的心理学量表。所有的参考资料都应该在后面参考资料部分里出现。如果用的是一个“家庭成长”的问卷，你应该描述测量方法是如何发展的。

过　程

这部分内容的写作规则非常简单。对所做的测试从头到尾准确地进行描述即

① 一种测试手指灵活性的实验仪器。——译者注

② 一种测试人感知觉的实验仪器。——译者注

可，就像按照配方制作蛋糕一样。此处，说明必须详细，使后人可以重复进行。说明最好标准化（也可以放在附录部分）。所有记忆实验里对被试进行培训所用图片和语言材料及其他的训练也应有详细说明。

我们试图要跳过枯燥乏味的方法描述部分，然而由于此处对于与读者沟通有着重要的意义，这里也和得分密切相关，因此请不要将此部分匆匆带过。最好你能使一个对心理学一无所知的人看了你的描述后，能按照你的步骤重复做下来，如果不能，说明你这部分的工作尚待完善。

结　果

描　述

在这一部分，对结果的描述同前面部分的写作风格的照应是相当重要的。你是在用语言告诉你的读者你发现了什么。表格和图可以对交流起到重要的补充作用。你不可能将单单的一个数据表作为你的"实验结果"。原始数据（如被试的原始分数）应该附在报告后面。所有的数据（平均数、标准差或者频次等）应呈现在同一表格上，而不要把每个被试的数据分别单独呈现。不需要对所有你能想到的数据的集中趋势（central tendency）进行一一描述。对统计数据的适当挑选会为你的报告加分。平均数、中数或者众数中的任何一个都可能是合适的，而无需将三者一一呈现。

所有表格，不管是这里出现的还是附录里出现的，都应该有完整的题头和标签。例如，表 14.2 所示中的实验结果的总结就不合规范。那些数字代表了什么测量值？表格需要有个题头，如"控制组在图片记忆时被试回忆出的正确的单词量"。如果呈现的结果是时间，要说明是以分还是秒为单位的；如果测的是距离，要说明距离单位是什么。表格的题头就是要告诉读者表中的数据是对什么内容的测量值。

表 14.2　信息不全的数据表——缺少题头或者对所测项目的说明

	图画组	控制组
平均数	12.4	8.3
标准差	1.5	1.1

或许你还需要为你的数据提供图，如条形图（bar chart）或者散点图等。同样，这些图也应有清楚的题头，坐标图的横轴和纵轴也应有相应的说明。如果你想用不同的颜色、阴影来表示不同种类的数据，那么你的图表也应配上相应的图例。不要为条形图配上类似于"第一组""第二组"这样的说明，应该使用描述性的说明，如"图片组"和"词语组"。

图表应在研究报告中有相关文字说明的部分出现，而不是出现在附录里。图表除了要有题头文字说明外还需要有数字编码（可以做参考用），数字编码要和文字题头放在一起。需要说明的是图表是有轮廓的图形，而表格则主要是为展示数据所用。（如表 14.2）

再者，不要滥用表格，不要用从不同角度对数据进行分析的表格来充抵你本身应做的工作。通常情况下，一个简单的实验只需要一个分析图来说明实验的主要效果。为什么读者对每个被试的单独得分的分栏图感兴趣呢？为什么对于同样的数

据读者喜欢看饼图、条形图和线图？为什么读者们想要看使用参数来说明的数据的特征呢？扪心自问"我的读者想要什么？"而不是"如何才能使报告看起来更加完美？"过多的图表不会为你的报告增色多少的。

不要相信电脑绘制的图表。如果不了解图表表达的意思，你画出的图表很可能看起来很蹩脚。即使电脑绘制的图表也应有题头和标签，如果坐标轴上显示的是"变量 0001"，那么这张表则几乎毫无意义。

结果的分析或"处理"

完成数据描述并对它们进行推断性统计检验对读者来说是件好事。阐述针对哪些数据用哪一种统计检验是比较合适的，并用第 10 章中所概括的决定因素类型与统计检验方法之间的对应关系来判断所选用的检验方法的实际应用。（其中可能还包括为什么用非参数检验而不用参数检验的具体细节）。

不要用"被证明的结果是……"来描述实验的结果。要清楚地阐明，如在结合图片记忆和仅仅复述的情况下，所得平均分的差异。我们可以用多种方法来证明最简单的数据表格。

对实验结果进行清楚的描述，如果没有进一步的统计分析，还可以将结果同适合的临界值进行比较。说明为何选用此临界值，样本容量或者自由度，样本数量以及在虚无假设下的显著性水平（如"$p<0.05$"）。有些计算机计算结果显示 $p=0.000$，这种情况其实是不可能发生的，对于可能发生的事件来说发生的概率是不可能为零的。这样的结果可以报告为 $p<0.01$ 或者 $p<0.001$。表 14.3 中列举了一些显著性水平的表述可能造成信息丢失的几个事例。对于本书中任何一个检验都有一个可接受这种实验结果的说明。

专栏 14.3　不完整的显著性描述

显著性描述	丢失的信息
"t 检验的结果表明差异显著"	t 检验一次只能进行一项差异性检验，此处指哪个差异？显著性在什么水平？自由度是多少？双侧检验还是单侧检验？
"两个变量相关度很大"	哪两个变量？正相关还是负相关？相关系数是多少？差异是否显著？如果显著是在哪个水平上显著？双侧检验还是单侧检验？
"两种情况下，在 0.01 水平上差异显著"	什么情况？单侧检验还是双侧检验？

说明虚无假设是被拒绝或接受。对结论进行描述：如对"图像情景比复述情景下被试的记忆唤醒水平要高的结论被证明是正确的"。

如果你想把检验的计算过程也放在你的报告里，那么只能放在附录中。如今很多计算都是由电脑或专用计算器完成的。其中应用的软件和计算的间接结果可以在附录中提及，除非涉及特殊的数据处理方法，否则一般是不用出现在报告中的。

讨 论

在此部分不要试图对你所发现的内容惜墨如金,然后把话题转回到与你课题相关的另外一篇文章上。通常来说,即使有新的研究或相关的背景理论,它们也只能占很小的篇幅。你应以你在引言中写到的东西为依据。这里你可以以研究报告中引言里提到的观点来讨论你的发现。由于你的研究需要,或者整体评论时,偶尔出现了新的参考文献也是可以的,但是这种情况发生的概率应保持尽可能地低。

对研究发现的概述

这部分要做的第一件事,就是将你的研究结果同你的假设和要验证的理论以及最初的研究目的联系起来。其次,他们的研究结果应该同理论背景相联系,并说明研究结果能够证明原来的理论,或者说明由于出现与事实相违背或模糊不清的发现,而要对理论进行修改。如果出现了未预料到的,或者"奇怪的"结果,如非常极端的分数,则应将其作为次要问题进行讨论。渐渐地,这样的奇怪结果可能将研究者引向新的研究方向。如果你有很好的理由也可以试着对这样的现象进行解释。

研究方法的评价

一个尽职尽责的研究者会对他的研究设计和使用的研究方法进行评价,并找出里面的错误和不足的地方。这可不是吹毛求疵,研究报告的读者们可能会反复阅读,并指出文章中作者没注意到的错误。作者可以通过提前对"为什么这样的不足不会影响实验结果"进行一场辩论的方法来避免这样的批判。评论重点依赖于研究的结果:

a. 如果得到的结果同预期相符,我们应该着重检查设计以避免干扰变量而犯 I 类错误。如果我们希望差异不显著(例如要否定早期发现的差异),那就要从可能会隐藏差异或联系的研究设计和研究程序等方法中寻找原因了。
b. 如果没有得到预期的显著差异,应该寻找材料的随机误差(random error)(那些取得理想结果的实验可能也要受这些因素的影响)。是不是研究设计、研究过程或者研究材料在有些方面是不符合要求的?同时,应该检查是否有任何无关变量在某一方面对结果产生影响,导致研究结果的不理想。

实验或调查中不可能事事完美。因此,通常不必要对温度或者背景噪声未进行有效控制而进行单独讨论,除非你有很好的理由认为这些变量会对研究结果产生很大影响。

修改建议及补充

大多数的研究会导致更多的研究。从现在的考虑出发,或许我们可以对研究的设计提出一些修改建议,这样做的目的在于:一是可以检验产生的临界点;二是对研究课题的新方向提出新的建议或者进行进一步的思索。

如果你发现自己执著于某事的时候,尽量避免这样下意识的反应:我们测试的被试数量还有待增加。这样的话通常出于对实验、样本本质以及研究目的误解的情

况下。我们在第2章中,已经对为什么有些情况下并不是样本越大就越合适这样的观点进行了解释。如果实验控制得很严格,你在两种情况下对30名被试进行了测试,那么你完全不需要更多的被试了,尤其是你已经得到了差异显著结果的时候。如果你坚持认为需要更多的被试,那你必须对为什么会有这样的想法进行解释。如果你认为应该调查一下性别差异的影响,或者被试应该包括各种不同文化背景的人们时,同样要说出你的理由。随后,对于作者假设的他/她与其读者来自何种文化背景这一问题必须陈述清楚。但是在这种情景中,最重要的问题是为什么:为什么要主张改进你以前的建议?除非你可以为自己找到一个进一步研究的良好理由,否则你最好不要因为想不到其他什么东西,单单为了讨论而胡乱地添加内容。

结　论

报告应以一个最终的评论结尾,但通常为了避免重复摘要部分的内容,最后一段的标题要省略。这里你要做的,就是对整体的研究发现、它与相关理论或模型之间的关系,以及对未来的暗示,做一个总结性的评论。不要通过类似于"发现一个新的效应"和"这一研究结果将使所有临床心理学家受益"等这样的话来吹嘘自己的研究发现。

参考文献

在你的引言和报告中,不管何处出现过你参照的曾经发表过的每个条目,都要有对应的一级或二级的参考文献。你报告中曾引用过许多不同的研究结果时,完成这样的列表尤其困难。如果这部分内容你有所忽略,或者完成质量很差,它最容易引起有名望的导师的怒火。通常你会发现有很多东西是不能被称做参考文献的,这可能会使你感到很困惑。参考文献中到底应该包含哪些内容呢?这里有个帮助你进行判断的黄金法则:

> 如果你的文章中对一些东西进行了直接的引用,那么把它算做参考文献;反之则不要放在参考文献内。

如果你文中写到"阿诺德(Arnold,2003)说过……",这就是一级参考文献。这些资料告诉我们阿诺德最初的研究论文或发表的时间,你文中的信息是直接出自阿诺德2003年的著作中的。然而,像许多报告里出现的情况那样,你从史密斯2004年的著作中得到了关于阿诺德写于2003年的著作的信息,那么只有史密斯2004年的著作才是你资料的来源,这一点一定要搞清楚。你可以在文中这样写"阿诺德(Smith(2004),2003)曾说过……"。这里你使用的便是二级文献(secondary reference),在你研究报告结尾的参考资料表里只需要出现(Smith,2004)就可以了。尽管有些系统要求两者都写明(如标准哈佛体系(SHS))。心理学家们如今普遍使用的是美国心理学协会(APA)规定的标准,参考如下:

> 在撰写报告时你参考(但是没有引用)的书籍不需要列入参考文献表中。如果你使用了格罗斯(Gross,2001)的资料来寻找某些相关资料而并没有在报告中引用任何格罗斯著作中的内容,那么不能把格罗斯列入参考资料中。严格来说,如果你重读鲍尔(Bower,1977)曾说过的话,你的参考

文献里该写为"(Bower,(Gross(2001),1997)"。而你上述的参考文献属于二级参考文献。在目录(Bibliography)中可能经常引用一些附加的内容,而心理学研究论文里则很少会出现这种情况。需要注意的是社会学家(sociologist)常用目录,而心理学家通常把同样的内容称做参考文献。

网络参考文献有很多陷阱。如果可能的话最好使用包括作者信息的内容,如果没有作者署名的话,你可要对这些信息小心对待了。其内容的可靠性是非常值得怀疑的。引用这类信息要给出它的作者、网址、你找到这些信息的日期以及页码等。

参考文献系统。大多数心理学家都采用同本书介绍相同的参考资料的写法,即由美国心理协会推荐的哈佛系统版本。其格式如下:

对于**书籍**:作者姓名,出处,(年份),**书名**. 出版地:出版商。

对于**论文**:作者姓名,出处,(年份),**题目**. **杂志**,**期号**,部分,页码。

需要注意的是,期刊论文的题目要使用斜体字(也可以使用下划线)并大写每个实词的首字母,而对于书籍来说,书名要使用斜体字,并只用对第一个单词的首字母大写即可。有些问题较为难办,如有些文章是出自于某人收集的文集、政府报告或者理学硕士生的论文等。但所有这些工作最重要的一个标准,就是让读者能容易地找到你所引用的文章原文。

附　录

附录一般包括:计算,对被试的指导语的细节内容,记忆词表,问卷等。这些内容的页码要续在正文的页码之后。不同课题的内容要用不同的页码和不同的附录(如"附录 1","附录 2"等)。

一般介绍

你的报告需要有一个对全文导航的页面,虽然有些导师不鼓励学生这样做。一个导航页可以帮助你很轻松地找到文章中你想要找的部分。标题页上通常标明所做的所有课题内容,目录页有助于读者较快地找到想看部分的页码。

关键词

摘要(abstract)
哈佛体系(Harvard system)
剽窃(plagiarism)
一级参考文献(primary reference)
理论概述(rationale)
二级参考文献(secondary reference)
情景故事(vignette)

14.3　学生实践报告的评论

下面你将看到两篇虚构的学生报告。第一篇报告的写作水平不高,因此如果要以此为模板,请仔细考虑并认真参考我们在旁边做的批注。我这样做的理由是如果我对一篇写的很完美的报告进行评论,新手们在报告撰写时所能见到的典型错误就

很少了。而一篇包括了所有可能出现错误的报告，又是无法阅读且没有任何意义的。同时，我还要强调的是，在我所举例的报告里没有拼写和语法错误，但这在学生的研究报告里却是经常出现的。对于有很多拼写错误的文章来说，导师会建议学生使用拼写检查器或者查字典。要求良好的拼写并非是注重繁文礼节。拼写问题严重的作品读起来很困难，如果在学生时代形成一个良好的拼写习惯，它可以使你在以后的工作申请中不至于落选，并且不会出现别人在不得不读你写的东西时，出现尴尬场面或者你会因此被惩罚的情况。不论你使用什么格式，请避免使用文本格式，你完全不需要对字符间距也如此精打细算。

鉴于我的强烈要求，本书还附有一篇遵守平均水平的良好的报告。我坚持这样做的原因，是它可以作为大家以后写作的模板。虽然学生报告遵守了一篇好的报告的所有要点，老师还是要强调的是，在某些特定的语境中，学生还是可能会出现某种错误。因此，大家可以将这篇范文看做是特定情景中的一个较好的例子，而不是可以满足所有要求的黄金模板。

一篇"平均水平"的报告可能只是一篇中等水平的文章，但是由于各种委员会所制定的标准之间也有所差异，因此我还是不提倡对它进行正式的评估。文中可能存在遗漏和模棱两可的东西，但不会存在绝对的错误。由于误导人的东西很多，我对下面的评论进行了整理：

√ 好的观点

× 错误，遗漏，模棱两可；整体上来说可能会积累造成整片报告的失分。

? 一个奇怪或者说模糊的观点，可能不会使文章失色但是如果重复则绝对会导致低分。使用的传统格式和语法等本身看起来并不糟糕，但积累效应导致对文章的感觉也不怎么好（这个取决于自己的研究水平）

假设前面提到的材料都已经包含在内了。为了便于参考，评论与报告将会齐头并进。

14.4 一个关于知道作者的性别是否会影响对一段文字的判断的实验[1]（一般报告）

摘 要

我们[2] 的实验设计是为了弄清楚，当人们阅读一段文字时，人们是否会对作者的性别进行推测。我们要求39名被试阅读同一篇文章，并告诉其中的一半（19人）被试作者是名男士，而告诉其他人，作者是名女士。我们通过将作者名字称做"约翰·凯利"为一个版本（男性作者），以"珍妮·凯利"为另一个版本[3]（女性作者）。出于对刻板印象的考虑，我们希望"珍妮·

1 ？"一个关于……的实验"是不需要的。题目应当简练："作者性别对文章评价的影响"。

2 ？传统的报告中通常使用**被动**语态。如"得出的理论是作者的性别会影响对作品本身的判断"。"39名被试被要求……"。

3 √ 自变量描述清晰。

凯利”组的被试认为文章质量差些[4]。实验结果表明差异不显著[5]，因此接受虚无假设。原因可能是选用的文章过于中性化，在技术类的文章中，女士的得分可能会更低，而在儿童读物方面的文章中，男士的得分更低。如果这一结果被证明有效，则说明自戈德伯格（Goldberg，1968）的研究至今，人们的态度发生了变化。[6]

4 × 没有对因变量下操作定义。“认为较差”应如何测量（后面我们发现使用的是评分）？

5 × 实验结果几乎没有报告。测试结果是什么，这些结果是怎么得到的？

6 √ 有简洁的总结。

引 言

人们评判别人时，通常是带着刻板印象的。我们用看待事物的方式来洞察人类。我们通过我们学习到的框架来看这个世界，我们并未看到事物的真实面目，而是由我们的印象对他加上了我们的期许和偏见。布鲁纳（Bruner，1957）曾说过“要超越信息提供的内容”[7]。我们用所谓的线索来解释我们看到的东西的确在那里。例如我们看到路上停着一辆汽车，在它的后面有一座山，后面的山看起来只有汽车的两倍高，但是我们可以通过山离这里的距离来估计山的实际高度；当我们给美丽的风景拍照时，我们通常会避开天空中架设的电话线，因为我们知道这些信息对我们并无什么重要作用。布鲁纳和他的朋友们以他们所做过的关于认知的实验来证明我们受情绪、动机和场景的影响，并以此提出他们对认知的新观点。在他们的一个实验中，他们证明：儿童的糖罐既可以用来放糖也可以用来放沙子[8]，实验结果是：儿童认为罐子里装了糖时看起来大。因此说我们的认知是受过去经验和期望影响的（Duckes and Bevan，1951）。[9]

7 √ 所引的短语使用了引号并且给出了作者，发表日期等细节信息（这些信息也应在报告最后呈现）。虽然只是引用了很少的内容但所引页码应该标注出来。

8 ？（可怜的孩子，你不会认为人们会允许心理学家对一些孩子做这样的事吧！）

9 ？ × 有点偏题。认知判断的影响因素与报告有一定的相关，但是报告应更多地同社会认知和刻板印象相联系。

阿希（1946，in Brewer and Crano（1994））[10]的实验中给一些被试看一些关于人的形容词，这些形容词除了“温暖”和“冷酷”两个词不一样，其他都是一样的。这些形容词会影响我们对这个人的判断。他们期望“温暖”和“冷酷”会产生不同的影响。这样的理论在凯利（Kelley，1950）

10 √ 恰当的二次引用。作者使用的信息是来自于布鲁尔和克莱诺但并不是材料的原出处。

的实验中也得到了证实，她给学生介绍一个"热情"或者"冷酷"的人，结果表明学生们更喜欢"热情"的那个。"热情"者看起来的确与"冷酷"者大相径庭。

性别差异简直是个神话[11]。坎杜瑞夫妇(Condry and Condry，1976)给人们看了一段关于一个九岁的小孩看到恐怖盒(Jack-in-the-box)的短片。如果告诉被试说短片里是个男孩，被试会认为小孩的反应是"愤怒"，如果告诉被试短片里的小孩是个女孩，被试则会认为小孩的反应是"害怕"。迪欧科斯(Deux，1977)对许多研究进行回顾总结发现，在执行完成不熟悉的任务时女性通常把她们的成就归因于幸运，而男性则说是他们的能力使他们取得了今天的成就。这说明男士和女士已经接受了社会上的刻板印象并使之伴随终身[12]。麦科比和杰克林1974[13]年的实验(Maccoby and Jacklin，1974)也证明了男性通常用独立的术语来形容自己(如智慧、雄心等)，而女性则多用社会性术语来描述自己(如合作、忠诚)。

一个叫做[14]戈德伯格(Goldberg，1968)的心理学家让一些女学生读了一段男士或女士的文章(学生自己的猜测)。说文章是男士所写的时，得分要更高些。我们所做的就是类似这样的实验[15]。如果性别刻板印象对我们的判断造成影响，我们就可以期望这样的结果：如果告诉被试一篇文章是位男士所写，那么这篇文章将比告诉被试为女士所写时评价好[16]。

实验方法[17]

实验设计

实验采用独立样本设计[18]。被试分成两组。自变量是作者的性别，因变量是被试判断文章的方法[19]。

11 ×!!! 这里提出了一个大而无法证明的假设，例如在阅读发展率上有些不同。陈述中需要诸如一些，很多或者例子和趋势等这样的限定词。

12 × 在一个具体的结果后出现了一个过大的假设，需要限定词。

13 × 这是一个事后研究的评论而非一个实验。

14 ? 无需像"一个叫……的心理学家"这样的句子，删掉它们

15 × 假设的跳跃性太大并且有些突兀。作者直接从清晰的背景描述，而没有当前研究的介绍和基本原理的论述，就直接跳到了理论假设。

16 × 假设含糊不清。研究预测叙述了自变量，但是还应该给出因变量"性质"和"重要性"的操作性定义。因此应该有两个预测，一个是性质方面的，一个是重要性方面的。

17 √ 方法部分的内容齐全，并有正确的标题。

18 √ 设计正确，并是个真实验。

19 × 没有明确的因变量。这里不需要完整地描述但是应该对测量下一个操作定义，类似于"内容是用给定的10点量表进行测量的"。其他的控制没有明确的提到。

被　试

从大学餐厅选取39名同学构成随机样本[20]。原本有20名在男性作者组,20名在女性作者组,但有一名男性作者组的被试缺失。被试中除了一名是其中一位同学的朋友外,剩余38名全是学生。

20　× 从餐厅选出的被试通常意义上不能称之为"随机"。如果是真正的随机要说明如何达到随机条件的。对被试的性别没有提及,性别同给定的课题相关度很大并且是这个研究的目标。学生的类型和来源没有说清楚。

材　料

选自卫报(*Guardian Weekend*)①的一篇托斯卡纳区②的旅行游记(见附录1)全文共908个字,打印在两张A4纸上。同时我们采用了一个关于品质和重要性方面的一个10分的评分表(rating sheet),使读者可以对所读文章进行评分(见附录2)[21,22]。评分表上有几个问题,回答这些问题会引起被试对作者的名字的注意[23]。

21　√ 实验材料的描述很清楚。

22　× 这里对精确的因变量的信息有所隐瞒。读者仍然不知道量表如何使用,量表中的10分是高分还是低分呢?

23　× 即使问题是虚构的也要注意提问的技巧,为了使被试注意到作者的性别应该做一个更加清楚的解释。需要说明事件发生的顺序。

实验过程

让被试坐下并使其放松,告诉他们这里不存在严重的欺骗,也没有会使他们看起来愚蠢的测试。指导语是:希望他们发表对某一事物的看法,他们的意见将和其他人的意见放在一起处理,结果回收是完全采用不记名的方式[24]。然后我们给他们做如下说明。所有的一切都以标准化的方式呈现[25]。

24　√ 伦理道德方面的考虑执行得很好。

25　? 表达模糊,报告的演示部分是否标准化,还是只把被试指导语发给大家?

> 我们希望你能读一下我们给你的文章。第一遍请快速阅读,第二遍仔细阅读。读完后请回答附在文章后面的表中的几个问题。请尽力回答好每一个问题,但请保证你是按顺序一一作答的[26]。

26　√ 这里使用了具体的说明。

奇数号的被试读的文章作者为"Jean Kelly",剩下的被试所读文章的作者名为"John Kelly"。有一个被试由于失误把文

① 英国的一种流行报纸。——译者注
② 意大利行政区。——译者注

章顺序搞反了[27]。

然后被试就留下来阅读文章，实验者可以问一些同阅读任务无关的问题，如是否需要打开灯或者是否需要关掉暖气。如果有被试问到关于阅读方面的问题时，主试的答案如下："你可以以你的理解尽可能地回答，我们可以在评论完成之后讨论"那个问题"，现在所有的参与者做的事情都是一样的，谢谢你的合作。"主试需要仔细观察确保实验说明能够按照正确的顺序进行。

27 ？被试的分配应该在设计部分或者被试部分说明，但是毕竟报告了错误还是做得很好的。

结 果

从两个实验组那里收集来的结果的原始材料见附录3。平均数和标准差的计算见表1[28]。

28 × 表格缺少题头。对于数值的意义如"6.7"没有进行说明。题头应该为内容和趣味性得分的平均数和标准差。

表1

	作者性别	
	女性	男性
内容		
平均数	6.7	6.3
标准差	1.5	2.3
趣味性		
平均数	4.3	5.2
标准差	1.1	1.3

你从表1[29]中可以看出：男性在内容方面的得分较低，但是在趣味性方面的得分较高。造成这种情况的原因可能是人们认为男人的写作内容通常比较有趣，而女人则更倾向于写作的精确性而且在语言和语法方面更加擅长[30]。

29 ？对读者应进行总体描述，不应使用人称。

30 × 这样的解释或者探讨应放在讨论部分。这里只需要报告研究的结果即可。

分 析[31]

对男性和女性在内容和趣味性方面得分平均数上的差异使用独立样本的 t 检验。t 检验属于参数检验，其前提假设是样本取自于常态分布的总体。同时变量的同质性(homogeneity)和测量的间隔水平也应当注意[32,33,34]。

内容维度检验的 t 值为 0.97，趣味性

31 × 不是所有的报告都要求写这样的标题。

32 × 不需要对你所使用的标准进行反复的说明。

33 √ 数据检验和 t 检验的标准都组织和描述得很好。

34 × 这里没有说明为何采用 t 检验，应该给出一个标准使读者清楚这些数据可以进行 t 检验。

检验的 t 值为 1.43。两者的差异均未达到显著水平，因此都接受虚无假设[35]。

35 × t 检验结果没有按照惯例报告。没有自由度和 p 值。检验是单侧还是双侧的也没有说明。

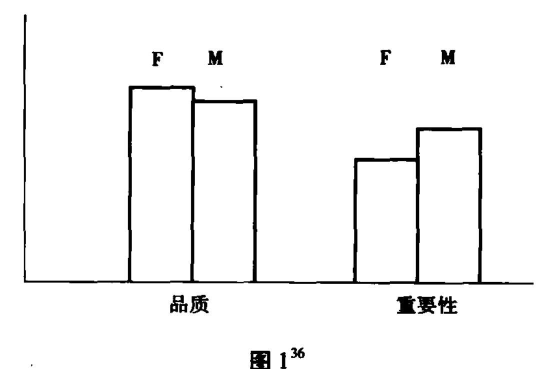


图 1[36]

36 × 图没有题头。"M"和"F"没有图例说明(尽管他们的意思大家可以明白，但是这里关键强调的是要规范清晰)；坐标的纵轴没有单位数值；图应该为丛集条形图而非直方图。文中没有任何地方对此图进行了相应的说明。

讨 论

如图 1 所示，男性作者组和女性作者组之间存在差异，但经检验未达到显著水平。造成这种结果的原因可能是两组结果之间本身有差异但是实验设计未能将这种差异表现出来[37]。也有可能是[38]当人们评价文章时作者的性别本来就不会对评论造成影响。如果上述假设成立那么我们早在 1968 年就可以并且已经对戈德伯格的实验结果提出了反驳。或者可能随着世事变化，人们不再像过去那样依作者的性别来判断文章的质量了。首先我们来看一下我们设计中可能出现的错误[39]。

37 √ 意识到可能会发生Ⅱ类错误。如果结果真实应该借鉴其他研究的结果对结果之间的矛盾进行说明。

38 ？语法问题！以这种方式开始一个新的句子是不合适的。

39 √ 要先处理Ⅱ类错误发生的可能性，如应批判地看待实验方法，找出未能证明预期效果的原因是什么。

我们要求被试回答我们所设计的问题，这样便可以在他们对文章进行评价之初就注意到作者的性别问题[40]。当我们事后重新考虑这一问题时认为或许应该让他们(至少其中的一部分)在阅读文章之前，先回答问题以使得他们在阅读文章时会注意作者的性别。这样的设计有可能得出不同的结论，以后有机会的话要做一个类似的研究[41]。

40 ？问题的设计应该早点说清楚，但是我们在前面的评论中已经提到过这个问题，这里不进行重复评价。

41 √ 对当前研究的结果和缺点提出修改建议。

虽然我们对被试的性别没有给予过多的关注，但很显然这可能造成差异。可能男性被试会对女性作者给出偏低的评价(downrate)[42]，反之亦然。在将来的研究中应当将被试组中的男士与女士分开[43]。所有的被试不会用同样的标准对我们提供的

42 ？这里用"评价(downrate)"一词是否恰当？查词典！

43 √ 好！对前面提到过的内容进行呼应，使我们可以进行对比评价。

量表进行评分,这是存在的另外一个问题。有的人觉得好时可能会评7分,但有的人却评9分。这个问题可以通过先训练被试对其他事物进行评分然后再同他们讨论量表评分的标准来解决这一问题[44]。同时我们可能需要更多的被试[45],被试可能会对实验目的进行猜测,因此可能会产生要求特征[46]。

实验所用的文章主题非常中性。戈德伯格使用的精选文章中,有些是传统的男性主题类文章,有些文章的主题则同女性联系更为密切。我们可以像米歇尔(Mischel)曾经做过的那样,使用汽车维护的一篇文章和幼儿护理的一篇文章看是否会得到不同的结果[47,48]。

也许被试会期望旅行作者是名男性。艾森克(Eysenck,2004)曾说过"视觉或听觉的特定信息增加了人们思考和行动的刻板印象的一致性"(P774)[49]。旅行作者是男性的刻板印象,可能使被试产生男性作者的文章更好的感觉,而事实却并非如此。然而,斯坦戈等人[50](Stangor et al,1992)发现刻板印象弱或者中等的被试对刻板印象不一致的信息的记忆要优于和刻板印象一致的信息的记忆。然而对于有着较强刻板印象的被试来说,对同刻板印象一致的信息的记忆效果要好于同刻板印象不一致的信息。这就是说,记忆过程只是加强了已经有较强刻板印象的人的刻板印象程度[51]。我们选取的被试都是学生,还没有形成很强的刻板印象,因此实验的结果就是有人给女性打分高(非一致信息),而有人给男性打分高,因此导致最后结果是没有差异的[52]。

参考文献[53]

Asch, S. E. (1946) Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 4, 258-90.

Brewer, M. B. and Crano, W. D. *Social psy-*

44 √ 这一点前面同样提到过,但是同 t 检验相关度不大,这能够被当做间隔水平么,或许有些偏颇。

45 × 为什么?除非有很好的理由否则应该完全避免这种凭直觉的观点。有相当一部分的被试无端地给出空洞的观点,对此我们完全可以摒弃。

46 ×?一个难点,被试是否可能进行猜测,这就可能存在着"要求特征"?如果答案是肯定的,那么报告中应该解释要求特征效果受到怀疑的原因和在哪方面应该受到怀疑。将被试的猜测作为要求特征是否可行?在独立样本设计里,我们要记住的一点是实验者知道自变量是什么,被试如何知道这点呢?他们为什么要怀疑其他的作者会是不同性别呢?这里需要考虑一下被试的观点。

47 √ 尽管研究需要引入更复杂的统计方法仍不失为一个很好的研究提议。

48 ×"米歇尔"没有出现相关的日期,也没有在参考文献中提到。

49 √ 承认引用了别人的话,并有明确的页码。

50 × 这样听起来可能不错,但是报告的惯例应当是:当你第一次提到一个出版物时,除非他有五个以上的名字否则不适用"等"这样的字眼。(文中实际上只有两个,因此需要把两个作者都写上,使用"等"是不恰当的)

51 × 警告评分者!遇到一个这么善变的学生时,这样的评价同报告的其他部分相比听起来过于学术化。多数有些许经验的评分者会发现这种变化,并会在书本上查找是否存在剽窃现象。事实上,这段材料直接来源于艾森克(Eysenck,2004:774)。这里只提供了少量的线索。对这样的报告来说这可以说是一个耻辱。参考文献中提到的应该是二级文献,艾森克应该被放入一级文献。

52 × 解释得不太好,最后一个观点有点模糊,应该有个整体性的结论而非这样生硬的结局。

53 √ 很好的参考文献。大多数都遵从了惯例并按字母顺序排列。然而,米歇尔也应该在这里出现。

chology. St Paul, MN: West Publishing Company.

Bruner, J. S. (1957) Going beyond the information given. *In Contemporary Approaches to Cognition: a symposium held at the University of Colorado. Cambridge*, MA: Harvard University Press.

Condry, J. and Condry, S. (1976) Sex differences: a study in the eye of the beholder. *Child Development*, 47, 812-819.

Deux, K. (1977) *The social psychology of sex roles*. In Wrightsman, L., *Social Psychology*. Monterey, CA: Brooks/Cole.

Dukes, W. F. and Bevan, W. (1951) Accentuation and response variability in the perception of personally relevant objects. *Journal of Personality*, 20, 457-465.

Eysenck, M. W. (2004) *Psychology: an international perspective*. Hove: Psychology Press.

Goldberg, P. (1968) Are women prejudiced against women?*Transaction*, April, 1968.

Kelley, H. H. (1950) The warm-cold variable in first impressions of people. *Journal of Personality*, 18,431-439.

Maccoby, E. E. and Jacklin, C. N. (1974) *The psychology of sex differences*. [54]

Stangor, C. and McMillan, D. (1992) Memory for expectance-congruent and expectancy-incongruent information; A review of the social and social-developmental literatures. *Psychological Bulletin*, 111, 42-61.

Rogers, W. S. (2003) *Social psychology: experimental and critical approaches*. Maidenhead: Oxford University Press. [55]

54 × 没有出现出版刊物和出版社。

55 × 它怎么能出现在这里呢？既没有按照字母顺序而且文中也没有提及它。可能是作者在准备写研究报告时参考过它，但它不属于参考资料。可以被放入“背景材料阅读”里，但是除非它为某些“一级”参考文献的出现提供了一些证据，否则没有必要提及它。

14.5 问题报告——更规范文章举例：作者的性别对书面文章评价的影响（好报告）

摘 要

为了研究性别角色刻板印象对写作技巧评价的影响，我们做了一个关于戈德伯格（Goldberg，1968）实验的验证性研究。实验要求25名被试阅读一篇文章，并告诉其中的12名被试作者是位男性，而剩下的13名被试则被告知作者是位女性。被告知是"女性作者"的文章在内容方面得分较"男性作者"的文章高，而在趣味性方面"女性作者"文章较"男性作者"文章得分较低，但两个方面的差异均未达到显著水平。之后对所评文章的中立性进行了讨论，如果文章本身已经被贴上性别标签（如技术文论或者儿童教育类文章），那么读者对文章的判断可能会随着作者性别的变化而有所不同。实验结果为"从戈德伯格的实验之后人们的社会态度发生了变化"这一观点提供了暂时性的支持。

引 言

正如我们对于物质世界的认知，受我们的主观以及认知的本质建构和解释的影响一样，我们对于人的认知也是如此。用布鲁纳（Bruner，1957）的话来说，我们对客观世界的认知构建是"超出所给定信息"的。

对于和性别相关的行为和特征的期望，会使我们对人们的认知产生很大的影响。康德利双氏（Condry and Condry，1976）曾做过一个研究，即给被试播放一个9个月的儿童看到恐怖盒的反应的短片。结果发现，如果告诉被试，短片中的儿童是个男孩时，其反应多被描述为"愤怒"；如果告诉被试，短片中的儿童是女孩时，其反应则被描述成"害怕"。这个实验说明，人们对其他人的评价可能不知不觉地受人们对性别认识的影响。作为行为表现经常被评价的学生，我们感兴趣的是对一个人的性别认知是否会影响我们对其文字作品的评价。在如今这样一个要求机会平等的大背景下，这个问题就显得愈发重要起来。一个人的能力是不应该受到其性别影响的。

性别偏见可谓屡见不鲜。戈德伯格（Goldberg，1968）的一个经典研究表明，女学生对看似是男性作者的文章的评分通常要高于看起来像女性作者的文章。米歇尔（Mischel，1974）的研究也表明，对男性领域话题的文章的评分，所有的被试给男性作者的评分都较高，而对于女性领域话题的文章的评分，所有的被试给女性作者的评分都较高。这种现象引起了人们对被试在进行评价时，是否会不自觉地用到关于作者性别的信息这一问题的关注。而作者的性别也可以作为实验的自变量。

在压力情景下和/或可用信息非常少时，刻板印象的知识存储会对人们的判断产生更大的影响。菲斯克和泰勒（Fiske and Taylor，1991）支持印象形成的**连续模型**认为人们在最初遇到或者听到某事的时候会对其产生一个简单的原始分类。如果我们不与其进行进一步的互动，人们将不会对信息进行进一步的分析。然而，如果我们需要与之进行进一步的互动或者进行评价时，我们才会继续寻找更多的信息。

如果有一个关于该分类的流行的、普遍的刻板印象，那么这个刻板印象将会继续奏效，直到一些与之冲突的信息使我们不得不对其进行重新分类（如，一个“硬实力”的商人却不肯为其职工修建育儿设施这样的事情）；或者我们的刻板印象被强化了（比如见到一个乏味并且自以为是的政客）。这样的例子不胜枚举，还包括了人们在匆忙时比悠闲时更容易以刻板印象作为判断的依据（Pratto and Bargh，1991）。纳尔逊、阿克和马尼斯（Nelson，Acker and Manis，1996）的研究表明，如果被试知道他们可以稍后进行评价，或者性别信息让被试产生挫折感的话，刻板印象对评价产生的印象就会有所降低。

按连续模型的预测，在戈德伯格的研究中，被试得到的信息太少（只有作者的名字和他们所写的文章）却要完成判断作者的个人能力，或者至少评判他们创作的文章这样的任务，这种情况下被试可能会受到性别信息的影响。历经30年，为了验证这种效应是否仍然存在，我们对他的实验进行了部分的验证，使用的材料是戈德伯格使用过的包括男性主导性文章和女性主导性文章中的一篇颇为“中性”的文章。戈德伯格所使用的被试都是女性，而我们的实验中男生和女生都有。我们要求被试对文章在“内容”和“趣味性”两个方面进行评分。如果当前流行的刻板印象在学生中仍然存在，我们希望被试对被告知是男性作品的文章的评分的平均数在内容和趣味性方面都应该高于被告知是女性作品的文章。

实验方法

实验设计

实验采用两组被试的独立样本设计。自变量为作者的性别，我们通过告诉一组被试文章的作者是男性，告诉另一组的被试文章的作者是女性的方式来操作自变量。通过设计一些问题来引起被试对作者性别（通过使用不同的姓名）的注意。因变量是文章在内容和趣味性（都采用10点量表）方面的得分。

被　试

从学生会中尽可能随机地挑选39名被试作为样本。其中，男性作者组中有12名男生和8名女生，女性作者组中有9名男生和10名女生。被试按照选择时的顺序交替分配到两种情况中。被试中除了一名是一位学生的朋友外其余均为学生。每种情况的原始被试人数相等，但后来有一位被试的结果缺失了。

材　料

使用的文章是选自《卫报·周末版》的一篇托斯卡纳区（意大利行政区）的旅行游记。原始文章提供在附录1中。全文共908个字，打印在两张A4纸上。材料有两个版本，一种作者姓名为“约翰·凯利”，另一种作者姓名为“珍妮·凯利”。我们采用一个以10分为最高分的10分量表（见附表2），使被试可以对文章的内容和趣味性进行评分。同时为使被试在对文章评价时已经对作者姓名、性别等信息有所注意，我们使用一些自编的问题（如文章的题目是什么，文章共几页等）。由于关于作者姓名的问题放在8个问题中间，我们认为这样可以避免被试通过这一实验程序猜

测到实验目的。对文章的内容和趣味性的评分在回答这些自编问题之后进行。

实验过程

主试要求每位被试坐下以使他们保持放松。告诉被试他们不会被欺骗,也没有被"测试"或被愚弄。让被试明白,研究者只是想知道他们对某事的看法,而且他们回答的结果将和他人的结果放在一起处理,回答的结果采取不记名的方式。按照如下指导语指导被试。所有的陈述和说明都需要标准化。

> **使用说明**
>
> 请阅读发给你的文章。第一遍请快速阅读,第二遍可以慢读。完成阅读后请回答附在文章后面的问题。请按照给定的顺序尽可能好地回答这些问题。

奇数号的被试发给女性作者"珍妮·凯利"版本的文章,剩下的被试则发给男性作者"约翰·凯利"版本的文章。有一次,我们由于失误把材料发反了。

然后被试开始阅读文章并回答问题。除了像被试要求开灯或者关掉暖气这样与文章无关的问题,主试不回答任何问题。被问到同阅读有关的问题时,标准答案是:"你可以以你的理解尽可能地回答,我们可以在完成之后讨论这个问题,现在所有的参与者做的事情都相同,谢谢你的合作。"主试需要仔细观察以确保实验说明按照正确的顺序进行。

结 果

从两个实验组那里收集来的资料结果的原始材料见附录3。统计结果的总结见表1。男性作者组(6.3)在内容上得分的平均数要低于女性组在此项上得分(6.7)的平均数;而在趣味性方面男性作者组得分的平均数则高于女性作者组得分的平均数(女性作者平均数 =4.3;男性作者平均数 =5.2)。男性作者在两个因变量上的得分尤其是在文章内容方面的离差均较高,(男性:sd = 2.3;女性:sd = 1.5)。平均数值见图1。

表1 在男性作者和女性作者情况下被试在文章内容和趣味性方面得分的平均数(标准差)

	文中作者性别	
	女 性	男 性
内容		
平均数	6.7(1.5)	6.3(2.3)
趣味性		
平均数	4.3(1.1)	5.2(1.3)

分 析

由于使用了10点计分的标准量表,容易让人觉得数据并不是来自真正的等级量表。检验男性作者和女性作者两种情况下文章内容和趣味性得分的差异时,使用非参数检验的曼-惠特尼法来分析降序排列的不相关数据。文章内容方面得分的平均数为女性作者6.2,男性作者6.1;在文章趣味性方面,女性作者为4.5,男性作者

为4.9。

在文章的内容方面，$U(N_a=20,N_b=19)=183, p>0.05$。在文章的趣味性方面，$U(N_a=20,N_b=19)=164, p>0.05$。（$N_a=20,N_b=19$ 时，临界值 $U=119$）。两个虚无假设都成立。

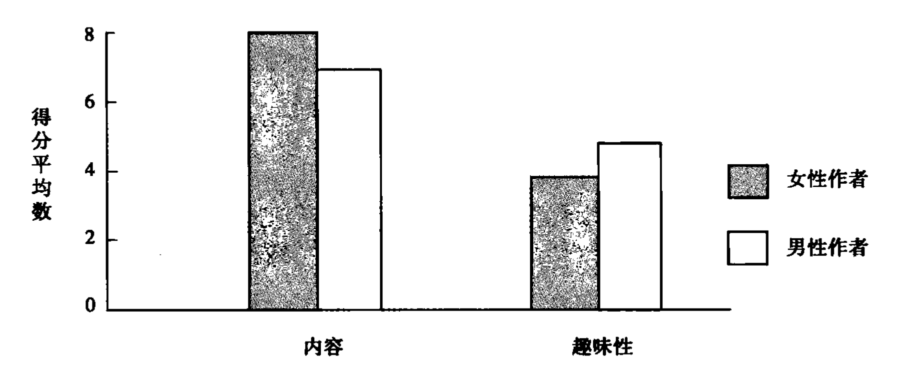


图1　男性和女性作者情况下内容和趣味性的平均分

讨　论

尽管在文章的内容和趣味性方面男性作者组和女性作者组存在小小的差异，推理分析的结果显示差异未达到显著水平。事实上，尽管男性作者在趣味性上得分的中位数高于女性作者组，但是女性作者组在文章的趣味性上得分的平均数要高于男性组。然而，对于使用的推理检验来说，数据的中位数相关度太高因此得出的差异更小。当前研究所能得出的唯一结论就是作者的性别对于文章的评价不会产生影响。对戈德伯格实验结果验证的失败可能是由于自他的实验以来人们对性别角色的刻板印象真的发生了改变。也有可能是普拉图、巴奇（Pratto and Bargh 1991）和纳尔逊等（Nelson et al., 1996）对戈德伯格效应作了大致比较。所谓的戈德伯格效应指的是在过去的30年里由于女权主义的出现和机会均等的社会环境使得社会性别刻板印象在公众面前变得极不明显。另一方面，也有可能是我们的设计没能测到已有的差异，我们应该看到可能造成这种结果的设计的缺陷在哪里。

我们要求被试回答一些设计好的问题是为了确定被试在对文章评分时已经注意到作者的性别问题了。事实上，这种做法并不一定能够确保被试的刻板印象在阅读和评价的过程中起到作用，而非仅仅在要求被试进行评价时才起作用。解决的办法是在被试阅读文章之前先提问设计好的问题，其中要包括像作者姓名这样的题目，以使被试在阅读的过程中注意到作者的性别。

戈德伯格的被试都是女大学生。可能是女性更容易对男性作者给出更高的分数，**反之亦然**。然而，在我们的研究中，由于我们的被试的性别不是单一的，因此由于女性被试对于男性被试产生的偏见可能被女性作者被试组中男性被试所产生的偏见抵消。我们可以对实验结果再次进行单独的分析，但是每类中可能只有很少的数据（例如女性作者组的男性的得分等）。正如米歇尔（Mischel, 1974）的研究所证明的一样，本实验中也有关于男性主义和女性主义的内容的交互作用。我们可以使用明确的有男性倾向和女性倾向的文章，如一篇关于汽车维护的文章，和一篇育儿方面的文章来重复这个研究（尽管这些课题现在已经收到如今典型的“男性”和“女

性”课题的挑战)。在这个实验中我们没有对文章内容的男性倾向或者女性倾向进行测试。我们可以通过要求其他被试在此维度上对文章评分或者通过增加一个不给作者姓名的控制组来研究男性组和女性组被试之间的差异,我们可以对这个无关变量进行部分控制。

鉴于我们的量表仅仅是为了研究的需要而编制的,所以我们没有证据表明所编量表可以按照标准化方式对所有的被试进行测试。为了表示“好”,有人可以用7分,也有人用9分。这个问题可以通过先训练被试对其他事物进行评分,然后再同他们讨论量表上的分值这样的方法使得量表评分更加标准。

这里很难看出要求特征对产生一个没有差异的结果起了什么作用。鉴于被试是第一次参加这样的实验,这些对实验目的毫不知情的被试是不可能猜出实验目的或者假设的实质的。然而,有些学生可能会在别处参加过类似的实验设计,而且通常情况下人们对于心理学的实验都有些心存疑虑。他们或许会怀疑作者性别这个自变量,但他们的怀疑可能仅限于基于种族和国籍而产生,比如对作者姓氏的假设的怀疑。解决的办法是检测一个非学生样本,和/或增加一个在任务报告讨论中关于怀疑的一个问题。

贝姆(Bem,1975)把性别刻板印象比喻为一件“紧身衣”(Gross,1996:696)并声称人们对“雌雄同体”态度的改变可能导致了社会的进步。或许,从我们的结果中,我们可以暂时这样认为:现在所发生的变化使人们在判断写作的质量时更少地把性别因素考虑在内了。也可能并没有发生什么变化,而是判断中关于性别刻板印象的假设的作用削弱了而已。

参考文献

假设这一部分没有出现任何问题!

附录一：术语解释

摘要(abstract)：在实践报告或研究文章开头出现的对假设、设计、方法、结果和结论的概括。

备择假设(H_1)(alternative hypothesis)(H_1)：假设一个效应的存在(例如，总体差异)。

匿名(anonymity)：使实验被试或来访者身份免于公开或避免遭到任何可能的无意泄露。

条形图(bar chart)：横轴代表类别变项，纵轴代表频数、平均数、百分数等的图表。

组间(between groups)：见"独立组"。

偏差样本(biased sample)：从大于或小于目标总体的子群体中进行抽样。

计算值(calculated value)：见"推断检验统计量"。

个案研究(case study)：深度研究所收集的个人或组群的数据。

人口普查(census)：对整体人口的调查。

集中趋势(central tendency)：描述一组数据的典型量或中间量的正式术语(如算术平均数、中位数、众数等——译者注)。

卡方检验(chi-square test)：对两个使用称名数据的类别变量的关联性进行检验。

组距(class interval)：直方图、类别或全距上的一个单位，在每一单位里都有频数呈现。

封闭式问题/项目(closed questions/ items)：在心理量表的选项中只能选择固定设置的答案，例如"是/否"。

编码系统(coding system)：通过给相似的行为观察实例赋予数值的量化的观察系统。

编码(单元)(coding (unit))：在定性数据中使用内容分析来确定项目类别。

同期组群(cohort)：为纵向研究或横截面研究确认的大样本人群，往往是同龄儿童。

结论(conclusion)：对研究结果的推断，即试验性地假设它们所出现并显示的效应。

同时效度(concurrent validity)：检验结果与同时施测的其他有效测量结果的一致性程度。

机密性(confidentiality)：避免公开涉及参加者或来访者的数据资料。

干扰变量(confounding variable)：一个没有被有效控制而产生干扰效应的变量。它通常在一个系统内，不是期望的自变量，却能对因变量产生影响。

结构效度(construct validity)：测验结果支持研究假设的程度，该研究假设是基于理论上的心理特质的变量。

内容分析(content analysis)：对质性材料的调查(尤其是文本)去发现"编码单元"(通常是单词、词组或主题)。分析通常聚焦于对频数的定量处理，但可以是一种纯粹的定性方法

内容效度(content validity)：测验能覆盖由专家评估产生的相关主题领域的程度。

控制条件/控制组(control condition/ group)：被用来作为基准衡量的组别或条件，用以对照实验组的表现进行评估。

方便样本(convenience sample)：抽取对测试而言极易获得的样本。

相关性(correlation)：两个变量之间的相关程度。

相关系数(correlation coefficient)：反映两个变量之间相关性程度的数值。

相关研究/相关设计(correlational study/ design)：通常是指在实验室外测量的非操纵变量间，一个变量与另一个变量的相关程度的研究。

平衡(counterbalancing)：一半的被试按照一个特定的顺序去做，而另一半的被试按照相反的顺序去做。这样做的目的是为了平衡在重复测量设计中可能出现的顺序效应。

效标效度(criterion validity)：测验所得分值能被用来对其他一些变量进行确切预测的程度。

临界值(critical value)：统计检验必须要达到拒绝

虚无假设的统计值(例如,t 必须 $\geqslant 2.086$, $df=20$, $p \leqslant 0.05$,双侧检验)。

跨文化研究(cross-cultural study):对两个或更多不同社会或社群/种族类别的比较研究。

横断研究(cross-sectional study):在一个时间内对几个不同组群进行测量的比较研究。

交叉表(cross-tabs table):一个变量的不同水平在另一个变量不同水平上的频率分布情况统计表。

累加频率(cumulative frequency):(表格或图表)的分布情况,以显示已出现的个体数量且包括当前的个体。

曲线关系(curvilinear relationship):两个变量间具有很低的相关系数 r,因为它们的关系不符合线性而是很好地符合曲线性。

数据集(dataset):在调查中已经收集好的数值集。

任务报告(debriefing):在实验前让被试简要了解情况,实验后完全告知被试研究的内容并让被试能够恢复到实验前的状态。

欺骗(deception):引导被试相信,有另一些因素而非真正的因变量参与或隐瞒了一些信息,导致调查现状的真实性被掩盖或扭曲。

需求特征(demand characteristics):研究中的线索以帮助被试了解对他们的期望是什么。

因变量(dependent variable):在实验中,被假设成受自变量变化影响的变量。

描述统计(descriptive statistics):描述数据库分布特征的统计方法。

离差分数(deviation score):一个值到平均数之间的距离。

定向假设(directional hypotheses):在总体中,提出差异方向性的假设。

语义分析(discourse analysis):对交互式语言的定性分析,例如人们根据背景和个人兴趣使用语言来建构他们所看到的世界。

离散数据(discrete data):只能分散在孤立单元上的数据(例如,孩子的数量)(在数轴上表现为若干孤立的点,相邻的两个取值间没有取值——译者注)。

离差(dispersion):反映一组数据的值在其平均数周围的分散程度。

依赖于分布的检验(distribution-dependent tests):更正规准确的术语是"参数检验"——见条目。

双盲实验(double blind):实验流程中无论是被试还是数据收集者都不知道被试将接受怎样的处理。

生态效度(ecological validity):研究效应在多大的程度上可以被推广到其他地方和条件下,特别是指从人工控制的环境(实验室)到自然环境的可推广程度。

实证的(empirical):在真实世界中基于对数据或对"事实"的观察。

启发(enlightenment):心理学学生逐渐意识到调查研究会经常涉及欺骗或其他的调查者"诡计"。

伦理(ethics):在应用心理学中对调查研究的被试,出版者,当事人的礼貌原则和专业责任。

评价顾虑(evaluation apprehension):被试对测试的担忧可能对结果产生的影响。

事件取样(event sampling):以观察事件为取样单元。

期望频数(expected frequencies):如果变量间不存在任何相关,即如果虚无假设为真,表格中期望的频数。

实验(experiment):在研究中,对一个自变量进行操纵,对一个因变量进行测量,而其他变量保持不变。

实验条件/实验组(experimental condition/group):接受自变量在不同水平下的处理的组别或条件,不同于控制组或控制条件。

实验假设(experimental hypothesis):见"备择假设"。

实验现实性(experimental realism):为吸引注意,使实验变得有趣的效应以补偿实验的人工性或"需求特征"。

实验者(experimenter):管理被试的实验程序和收集实验数据的人。

实验者期望(experimenter expectancy):实验者有

关实验内容的知识影响研究结果的倾向性。

实验者信度(experimenter reliability):由两个或更多实验者操作结果的相关程度。

外部信度(external reliability):检验的稳定性,在相同样本上重复测验产生相同结果的倾向性。

外部效度(external validity):实验结果在跨人群、跨地域、跨时间上的概括化程度。

无关变量(extraneous variable):除了因变量和影响因变量的自变量外,其他的一切变量,它也许很难被顾及或被控制。

表面效度(face validity):研究者主观上觉得测验的有效性程度(但不能替代客观上的真正效度——译者注)。

现场研究/现场实验(field study/experiment):调查研究或实验的数据采集地点不是在被试通常所在的实验室。

结果(findings):对研究中所收集的数据进行归纳和分析,但并没有进一步的推测以从中得出什么结论。

拟合度(fit):定性的原始数据与相应解释之间的吻合程度(用建构的预测模型预测出的结果与实际结果的吻合程度。——译者注)。

0.05 水平($p \leq 0.05$)**5% level**($p \leq 0.05$)):传统的显著性水平。

焦点团体(focus groups):被要求就具体某个事项或主题展开讨论的团体。

频数(frequency):对某现象或某事件发生的计数。

频数分布(frequency distribution):在不同类别中频数的分布。

推广(generalizability):能够从样本推广到总体(样本所来自的总体)的程度。

拟合优度检验(goodness-of-fit test):频数的分布与理论模式是否有显著性差异的检验。

扎根理论(grounded theory):理论驱动下对定性数据的分析,并从数据中产生了理论模型。但在数据收集以前,并未强加理论模型。

组差研究(group-difference study):事后研究以比较两个截然不同的群体所存在的变量的测量结果,例如:内向或外向。

哈佛系统(harvard system):共同的学术参考系统。

霍桑效应(hawthorne effect):人在得知被他人观察时所产生的行为效应。

直方图(histogram):将连续的数据分解为具有比例常数间距的图表。

历史效度(historical validity):研究结果在不同历史时期的可概括化的程度。

方差齐性(homogeneity of variance):样本间或总体间的方差是否存在显著性差异。

假设(hypothesis):精确陈述变量间关系的假设,这些变量来自同一理论,变量间的关系可以得到具体的检验。

假设-演绎方法(hypothetic-deductive method):记录所观察的一种方法,从这些理论中进一步发展解释性理论和检验假设。

独立组/独立样本/独立测量(independent groups/samples/measures):每一组被试只接受自变量的一个水平的处理的实验设计。

自变量 independent variable):实验者在实验中操纵的、被假定是对因变量产生直接影响的变量。

推断统计(inferential statistics):在显著性检验过程,通过样本推断样本所来自的总体。

推断检验(inferential test):一种可以帮助我们从对样本的检验中推测总体效应的统计检验。

知情同意(informed consent):同意让研究的被试全面了解研究的背景(但并非是所有细节)以及被试的权利。

内部信度(internal reliability):以每个项目上被试得分的方差与量表所有项目的方差的相关性来衡量量表信度的方法。

内部效度(internal validity):研究所发现的效应在多大程度上被视为是真实的且由被操控的自变量造成的。

观察者内部信度(inter-observer reliability):见"评分者信度"。

诠释现象学分析(interpretive phenomenological analysis):尽可能贴近个体自身的视角;尝试描述他们的经历,但这种解释方法仍被视为研究者对研究对象产生了解释性影响。

评分者信度(inter-rater reliability):评分者在评分或编码上的一致性程度。

等距水平(interval level):量表上的每一个单位代表变量测量中的一个相等的变化水平。

干预(intervention):超出具体的研究背景来改变人们的生活,尤其是给人们创造利益。

访谈法(interview method):使用面谈来收集数据的方法,通常是一对一。

调查者(investigator):对调查项目全权负责的人。

调查者效应(investigator effect):由于调查者(行为或期望)的原因对结果产生不期望获得的效应。

非志愿参与(involuntary participation):没有获得本人同意或在对研究毫不知情的情况下参与研究的情形。

项目分析(item analysis):对一项测验的项目内部一致性的评估方法。

杂志(journal):定期出版的刊物里刊登研究中有新发现的文章。这是科学学科(包括心理学)不断发展进步的基础。

实验室研究/实验室实验(laboratory study/experiment):在研究者自己的实验室里开展实验的研究。

测量水平(level of measurement):对数据进行分类和测量的水平。

水平(自变量的)(level (of the IV)):由其中的一个条件或组别(水平),与其他的条件和组别一起构成的自变量。

李克特量表(Likert scale):在量表上使用对表述作不同反应的刻度,通常是从“强烈反对”到“强烈同意”。

纵向研究(longitudinal study):对一个个体或组别进行一段相当长时期的比较研究(可能包括控制组)。

曼-惠特尼 U 检验(Mann-Whitney U test):对两个独立样本数据进行秩和差异性检验。

配对设计(matched pairs):实验设计中每一个组别或条件下的一位被试都与另一个组别或条件下的一位被试在特定变量水平上配对。

平均数(mean):通过对所有数的相加并除以数集中的个数来测量集中趋势。

平均差(mean deviation):对离差的测量,所有绝对离差的平均值。

中位数(median):数集中的中位量。

众数(mode):数集中出现频率最高的值。

世俗现实性(mundane realism):实验设计的特点类似于日常生活,但未必是目前正在从事的情形。

自然实验(natural experiment):对在自然状态下自变量发生的事件进行实验。

自然观察(naturalistic observation):以在自然生活情景里发生的自然行为为观察对象而开展的观察设计。

消极案例分析(negative-case analysis):发现案例不符合建构的解释,就继续分析直到被某种理论解释。这是一种运用于扎根理论的方法。

负相关(negative correlation):当一个变量的值增加时与此相关的另一个变量的值呈递减趋势。

负偏态(negative skew):对分布情况的一种描述,它包含了低值端的一个更长的单侧分布。

称名水平(nominal level):数据仅反映类别的发生频率,即使需要使用数字,也仅是用于标识类别。

非定向假设(non-directional hypothesis):对总体不具有差异方向性的假设。

非等组(non-equivalent groups):在独立样本设计中,每种条件下的被试未必等价并由此可能影响实验结果的问题。这是被试变量的问题。

非参数检验(non-parametric test):在一个未知分布中,无法作参数估计的显著性检验。

正态分布(normal distribution):连续的分布,在中点处呈对称的钟形图。

虚无假设(H_0)(null hypothesis(H_0)):样本所属的各总体间(被检验总体与假设总体的参数间——译者注)无差异的假设。

观察设计(observational design):该研究设计中的观察是收集数据的技术手段,这种行为观察是相当不受约束的。

观察研究(observational study):通过观察和记录行为来收集数据的研究。

观察技术(observational technique):在有约束和结构性的设置中对观察和编码的使用。

观察频数(observational frequencies):在调查研究中使用分类变量所获得的频数。

观察者偏差(observer bias):仅是由于观察者特质导致的对观察记录效度的影响。

实际(或计算)值(obtained (or "calculated") value):见"检验统计量"。

单侧检验(one-tailed test):如果备择假设是方向性的,则可使用单侧检验(相反方向上的结果可忽略,即使该结果显示具有显著性差异)。

开放式问题/开放式项目(open questions/items):心理量表中的项目可以用自由、不受限制的人类语言来回答。

操作性定义(operational definition):以变量的具体测量步骤来定义变量。

机会样本(opportunity sample):抽取对测试而言极易获得的样本。

顺序效应(order effect):经历一种处理后经历另一种处理引起的与处理的顺序有关的干扰效应(例如,练习、学习或疲倦)。

顺序水平(ordinal level):数据(分数)按等级顺序进行排列的量表。

专门小组(panel):为了评估观点,可征询意见的分层组。

参数(parameter):对总体的统计测量(例如,总体平均数)。

参数检验(parametric test):相对强大的显著性检验,用于对总体参数的推断或估计,数据的检验因此符合某种特定的假设,有"依赖于分布的检验"之称。

被试(participant):在研究中作为被研究的人员。

被试期望(participant expectancy):被试认为应该发生的事情而产生的期望效应对研究结果的影响。

参与观察(participant observation):在分组观察中,主试参与活动或扮演角色的一种观察。

被试变量(participant variable):在不同实验组中反映人的差异性比例的变量,它可能会干扰实验结果。

被试核实(participant verification):被试对定性研究者最终解释的认同。

皮尔逊积差相关(Pearson product-moment correlation):相关性的参数计量。

预实验/预研究/预(pilot study/pilot trials/piloting):在一个小样本中进行初步的量表研究或试验,对可能遇到问题的推测,并评估一个重要的研究可遵循的特性。

安慰剂(placebo):给被试貌似真实却缺乏重要影响因子的治疗。如用去咖啡因的饮料替代咖啡因,检测自变量单独处理时的心理效应。

剽窃(plagiarism):提交的个人论著实际上是他人的。

取悦实验者(pleasing the experimenter):被试推测主试所希望看到什么,以此来表现他们的行为的倾向性。

时点取样(point sampling):以设置时间点的方式来进行观察,例如,每30秒间隔结束时。

总体(population):从中可以抽取样本的所有可能的个体总和。

总体效度(population validity):研究结果可以被推广到总体的其余部分或其他总体的程度。

正相关(positive correlation):这种相关的情况是:当一个变量的值增加时与此相关的另一个变量的值呈递增趋势。

正偏态(positive skew):对分布情况的一种描述,它包含了高值端的一个更长的单侧分布

实证主义(positivism):方法论上的一种观念,认

为现象可通过观察到的事实加以还原,并可测量。如果无法测量,则概念就不能成为科学的一个部分

事后研究(post-facto study):对已存在于人群中的变量进行测量来获得差异性或相关性的研究。

预测效度(predictive validity):测试分数对未来行为或态度的预测程度。

前测(pre-testing):实验前对被试的测量,用以平衡或比较组别情况。

概率(probability):一种针对基于随机性的"机会"事件的数值计量。

一级参考(primary reference):作者真实阅读过的资料来源——见"二级参考"。

投射测验(projective tests):基于精神分析理论的测试,尝试获得由潜意识欲望和冲突所导致的反应。

心理学构想(psychological construct):假设存在的心理现象,可用于解释观察到的行为。

心理量表(psychological scale):以纸笔为工具的测量设计,用来测量一种或多种心理结构。

心理测验(psychometric test):尝试量化心理变量的测验,例如:技能测验、能力测验、性格测验等。

心理测量学(psychometrics):建构心理测验的学科。

定性数据(qualitative data):以意义的原始形式(例如,语言、文本)保存的数据,而非量化的数据。

定性研究(qualitative research):主要收集定性数据的研究方法。

定量(quantity):对某一现象进行量化方式的测量。

定量数据(quantitative data):以数字形式呈现测量结果的数据。

四分位数(quartile):分别是 P25,P50,P75 三个百分位数。

准实验(quasi-experiment):主试没有完全控制住核心变量的实验。

准等距量表(quasi-interval scale):看似等距的量表,但实际上每个间距未必能测量到结构上相等的数量。

问卷(questionnaire):使用问题的提问工具。大多数心理量表的测量是不问问题的,而是呈现陈述,要求被试表达同意或不同意的程度。

定额样本(quota sample):以子群(层次)在目标总体中的比例来决定该子群的抽样比例,而并非是随机抽样。

随机分配(random allocation):以随机的方式,在实验中把被试放人不同的处理中。

(简单)随机样本((simple) random sample):样本的抽取方式使得目标总体中的每个个体都有一个被选中的平等机会,以及可以得出所有可能性的样本组合。

随机化(randomization):把各种试验或刺激通过随机序列放入到实验中,事前的任何预测都不可能。

全距(range):对离散情况的测量:从最高值到最低值。

评分者(rater):根据标准量表对访谈或其他内容进行评估的人。

等比水平(ratio level):由于绝对零的存在使得量表上的比例具有倍数意义的区间水平量表。

理论概述(rationale):在对前人研究回顾的基础上,出现在一份实践报告的前言中的观点:即证明为什么要开展此项研究的原因。

原始数据(raw data):收集在一项研究中未经处理的数据,例如,被试的原始得分。

反应性(reactivity):对被试意识的研究,这种意识可能导致他们相应行为的改变。

反思(reflexivity):研究者承认自己的知识、经验和态度会影响他们如何分析、建构、呈现研究结果。是质性研究的特点。

相关设计(related design):被试在一种处理下的得分与其他处理下的得分相匹配的实验设计,例如重复测量设计和配对组设计。

相关 t 检验(related t test):在区间水平上,对一组相关数据的参数差异性进行检验。

信度(reliability):测量结果或效应的一致性程度。

重复测量(repeated measures):每个被试都参加自变量的所有水平实验。

重复(replication):重复一项已经完成的研究。

研究设计(research design):确定数据采集的结构和研究的策略。

研究预测(research prediction):通过分析研究获得的数据,用精确的术语预测变量间关系。

研究问题(research question):研究者努力想在调查中获得回答的问题。

调查对象(respondent):回答心理量表、问卷或调查的人。

反应定势(response set):人们习惯性应答"同意测试选项"的倾向性。

撤回的权利(right to withdraw):被试有权从实验中撤出自己或已被采集的数据。

样本(sample):为完成一项研究,从总体中抽取出的小组。

抽样偏差(sampling bias):样本中的一些类别具有过代表性或欠代表性的系统倾向。

抽样误差(sampling error):出现在假设中的错误:认为总体参数与样本统计量是相同的。

饱和度(saturation):达到饱和点(特别在扎根理论中),此时额外的数据只作出轻微的贡献,并不能改变已出现的类别和主题框架。

散点图/分布图(scattergraph/scattergram):在一个二维图上显示坐标值位置的图表。

科学方法(scientific method):使用推测和演绎对经验世界进行研究的一般方法。

二级参考(secondary reference):作者提到的论著未阅,但在其他刊物上读到。如作者提到 Smith (1999)在 Bolton(2001)中读到:Bolton 为原始参考,而 Smith 为二级参考。

选择偏差(selection bias):见"抽样偏差"。

自我选择样本(self-selecting sample):自我选择是否参与研究并成为被试。

语义分化(semantic differential):量表对某个事物意义的测量方式是通过将被试的选择置于若干个两极形容词的两极间。

半结构访谈(semi-structured interview):虽然带着预先设置的主题清单,但仍然努力使得谈话显得"自然",采访者只是"带着耳朵"来倾听受访者。

(二项)符号检验((binomial) sign test):对相关样本的类别数据或称名数据的差异性进行检验。

显著性(significance):基于在虚无假设下对事件概率计算的基础上,是否应该保留或拒绝虚无假设的决断。

显著性水平(significance level):同意拒绝虚无假设的概率水平。如果在虚无假设下,所得结果的概率低于设置水平,就拒绝虚无假设。

显著性检验(significance test):见"推断检验"。

显著差异(significant difference):如果虚无假设能够成立,这种差异就不大可能发生。

单盲(single blind):实验流程中被试不知道他们所接受的是怎样的实验(例如,他们处于怎样的条件中)。

偏态分布(skewed distribution):分布情况是一侧比另一侧有更多的数值。

滚雪球样本(snowball sample):通过使用来自早期参与实验的关键被试的信息来获得更多的样本(先抽取少量的样本,然后通过样本信息以滚雪球的方式扩大样本。——译者注)

社会期望(social desirability):研究被试由于想要"看上去很棒"的结果从而提供社会认可的答案的心理倾向性。

斯皮尔曼等级相关(Spearman's rho (ρ) correlation):非参数的相关计量。

折半信度(split-half reliability):一项测验被拆分成相等的两个部分,这两部分得分的相关性。

标准差(standard deviation):对离散程度的测量,是离均差平方和除以 $N-1$ 后的方根。

标准分(Z 分数)(standard score (Z score)):一个特定分数距离样本平均数有几个标准差。

标准化(standardization):根据心理测试的对象,

对总体设置测量标准。

标准化指令/标准化程序(standardized instructions/ procedures)：对所有被试的测验或测量都采取完全相同的正式的操作规程和指令。

统计量(statistic)：对样本的统计测量（例如，平均数）。

分层样本(stratified sample)：子群在样本中所占的比例与该子群在目标总体中所占的比重一致，而每个子群中的样本是随机抽取的。

结构访谈(structured interview)：对所提问题的次序和内容极少发生变化的访谈。

结构化观察/系统性观察(structured/systematic observation)：使用明确定义的编码系统进行数据记录的观察。

调查法(survey)：对大样本进行相对结构化的询问。

系统抽样(systematic sample)：如果起点是从"*n*"点开始随机抽取的话，就要从目标总体的名单上选取每第 *n* 个个体。

检验统计量(test statistic)：在推断检验中获得的统计量（例如，$t = 6.54$）。

主题分析(thematic analysis)：使用定性数据来检验假设，通过这种方法，理论可以用于分析，但是这种分析是以事例的含义而不是定量数据作为依据的。

理论(theory)：在真实世界中事件之间如何发生关联的通常模式，是对什么引起了什么的推测。

效度威胁(threats to validity)：研究设计或研究方法的任何一个方面均有可能削弱了实际效应的显现或可能使一个实际效应的存在变得模糊。

时间取样(time sampling)：在设定的区间和时间长度内进行的观察。

处理(treatment)：操控变量来观察它们是否对行为产生影响，与控制条件下的行为比较，自变量水平操控对行为产生影响。

三角校正(triangulation)：至少对同一事物的两种观点或解释进行比较（例如，事件、行为、行动等）。

双侧检验(two-tailed test)：如果备择假设是非方向性的，那么就必须使用这种检验方法。

Ⅰ型错误(type Ⅰ error)：当虚无假设为真而拒绝它所犯的错误。

Ⅱ型错误(type Ⅱ error)：当虚无假设为假而接受它所犯的错误。

独立设计(unrelated design)：个体在一种条件下的得分无法与在其他任何条件下的得分相匹配的设计。

独立样本 *t* 检验(unrelated *t* test)：在区间水平上，对两组独立数据的参数差异性进行检验。

效度(validity)：检验所达到测量意图的程度。

变量(variable)：对变化的现象进行确切的量化或赋予分类值。

方差(variance)：对离散的测量，标准差的平方。

情景故事(vignette)：在调查研究中给被试描述某一场景的小短文，它经常是仅有某一特性会发生变化，用来充当自变量。

视觉模拟量表(visual analogue scale)：被试以图表的形式标明自己在某个两极维度上的位置。

自愿者样本(volunteer sample)：研究中由自愿者组成的样本。

威尔科克逊配对秩次检验(Wilcoxon matched pairs test)：两组相关数据的秩次水平差异性检验。

撤回权(withdraw)：被试拥有任何时候都可以从心理学研究中撤出的权利。

组内(within groups)：见"重复测量"。

***Z* 分数**(*Z* score)：见"标准分"。

附录二:数据表*

表1 标准正态分布表(节选)

Z	0 Z	0 Z	Z	0 Z	0 Z	Z	0 Z	0 Z
0.00	0.0000	0.5000	1.95	0.4744	0.0256	2.55	0.4946	0.0054
0.01	0.0040	0.4960	1.96	0.4750	0.0250	2.56	0.4948	0.0052
0.02	0.0080	0.4920	1.97	0.4756	0.0244	2.57	0.4949	0.0051
0.03	0.0120	0.4880	1.98	0.4761	0.0239	2.58	0.4951	0.0049
0.04	0.0160	0.4840	1.99	0.4767	0.0233	2.59	0.4952	0.0048
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表中三大栏中的每栏左侧一列显示的是 Z 值。中间一列显示的是平均数 0 和 Z 值之间的面积。右侧一列显示的是整个正态分布剩余部分到 Z 值之间的面积。正态分布下的整个面积是 1 个单位,所显示的数值是占整个面积的百分比。这些也是在所涉及区域内发现某个数值的概率。转换百分数时,所有面积值乘以 100。在 -Z 和 +Z 之间的面积,两倍于中间一列所显示的数值。

SOURCE:R. P. Runyon and A. Haber, *Fundamentals of Behavioral Statistics* (1976) 3rd edition. Reading, Mass.: McGraw-Hill, Inc. Used with permission. Artwork from *Fundamental Statisfics for Psychology*, (2nd edition) by R. B. McCall © (1975). Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www. thomsonrights. com. Fax 800 730-2215.

* 译本根据方便读者学习的原则对原书的数据表部分做了删减处理。读者若欲查询完整的数据表信息,请登录本书封底提供的链接地址。

表2 t检验的临界值

	单侧检验的显著性水平			
	0.05	0.025	0.01	0.005
	双侧检验的显著性水平			
自由度	0.10	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

要达到各水平上的显著性水平，计算出的 T 值必须等于或超过表中的（临界）数值。

SOURCE: Abridged from R. A. Fisher and F. Yates (1974) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. Pearson Education Ltd.

表 3　在威尔科克逊符号排序检验中的 t 临界值

	显著性水平			
	单侧检验			
	0.05	0.025	0.01	0.001
	双侧检验			
样本量	0.1	0.05	0.02	0.002
$N=5$	$T \leqslant 0$			
6	2	0		
7	3	2	0	
8	5	3	1	
9	8	5	3	
10	11	8	5	0
11	13	10	7	1
12	17	13	9	2
13	21	17	12	4
14	25	21	15	6
15	30	25	19	8
16	35	29	23	11
17	41	34	27	14
18	47	40	32	18
19	53	46	37	21
20	60	52	43	26
21	67	58	49	30
22	75	65	55	35
23	83	73	62	40
24	91	81	69	45
25	100	89	76	51
26	110	98	84	58
27	119	107	92	64
28	130	116	101	71
29	141	125	111	78
30	151	137	120	86
31	163	147	130	94
32	175	159	140	103
33	187	170	151	112

要达到所示各水平上的显著性水平，计算出的 T 值必须等于或小于表中的（临界）数值。

SOURCE: Adapted from R. Meddis (1975) *Statistical Handbook for Non-Statisticians*, McGarw Hill, London, with the kind permission of the author and publishers.

表 4　在单侧检验 0.025 显著性水平上或双侧检验 0.05 显著性水平 * 上的 U 临界值(曼-惠特尼 U 检验)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	0	0	0	0	1	1	1	1	1	2	2	2	2
3	—	—	—	—	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	—	—	—	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	—	—	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	—	—	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	—	—	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	—	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	—	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	—	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	—	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	—	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	—	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	—	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	—	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	—	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	—	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	—	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	—	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	—	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

* 表中短横表示在所陈述的显著性水平上没有决策的可能性。

对于任何一个 n_1 和 n_2,如果观察到的 U 值等于或小于表上所显示的与显著性水平相应的(临界)数值,那么它都达到显著性水平。

SOURCE:R. P. Runyon and A. Haber (1976) *Fundamentals of Behavioral Statistics*,3rd edition. Reading, Mass. :McGraw-Hill,Inc. with kind permission of the publisher.

表 5 χ^2 检验的临界值

	单侧检验的显著性水平					
	0.10	0.05	0.025	0.01	0.005	0.0005
	双侧检验的显著性水平					
df	0.20	0.10	0.05	0.02	0.01	0.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.60	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.80	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.00	28.26	30.58	37.70
16	20.46	23.54	26.30	29.63	32.00	39.29
17	21.62	24.77	27.59	31.00	33.41	40.75
18	22.76	25.99	28.87	32.35	34.80	42.31
19	23.90	27.20	30.14	33.69	36.19	43.82
20	25.04	28.41	31.41	35.02	37.57	45.32
21	26.17	29.62	32.67	36.34	38.93	46.80
22	27.30	30.81	33.92	37.66	40.29	48.27
23	28.43	32.1	35.17	38.97	41.64	49.73
24	29.55	33.20	36.42	40.27	42.98	51.18
25	30.68	34.38	37.65	41.57	44.31	52.62
26	31.80	35.56	38.88	42.86	45.64	54.05
27	32.91	36.74	40.11	44.14	46.96	55.48
28	34.03	37.92	41.34	45.42	48.28	56.89
29	35.14	39.09	42.69	49.69	49.59	58.30
30	36.25	40.26	43.77	47.96	50.89	59.70
32	38.47	42.59	46.19	50.49	53.49	62.49
34	40.68	44.90	48.60	53.00	56.06	65.25
36	42.88	47.21	51.00	55.49	58.62	67.99
38	45.08	49.51	53.38	57.97	61.16	70.70
40	47.27	51.81	55.76	60.44	63.69	73.40
44	51.64	56.37	60.48	65.34	68.71	78.75
48	55.99	60.91	65.17	70.20	73.68	84.04
52	60.33	65.42	69.83	75.02	78.62	89.27
56	64.66	69.92	74.47	79.82	83.51	94.46
60	68.97	74.40	79.08	84.58	88.38	99.61

计算出的χ^2 值必须要等于或大于表中的临界值，结果栏在该水平上显著。

SOURCE：R. A. Fisher and F. Yates(1974) *Statistical Tables for Biological. Agricultural and Medical Research*, 6th edition. Pearson Education Ltd.

表 6 二项符号检验的临界值

N	0.05	0.025	0.01	0.005	0.0005
	单侧检验的显著性水平				
	双侧检验的显著性水平				
	0.10	0.05	0.02	0.01	0.001
5	0	—	—	—	—
6	0	0	—	—	—
7	0	0	0	—	—
8	1	0	0	0	—
9	1	1	0	0	—
10	1	1	0	0	—
11	2	1	1	0	0
12	2	2	1	1	0
13	3	2	1	1	0
14	3	2	2	1	0
15	3	3	2	2	1
16	4	3	2	2	1
17	4	4	3	2	1
18	5	4	3	3	1
19	5	4	4	3	2
20	5	5	4	3	2
25	7	7	6	5	4
30	10	9	8	7	5
35	12	11	10	9	7

要达到各水平上的显著性水平，计算出的 S 值必须等于或小于表中的临界值。

SOURCE：F. Clegg, *Simple Statistics*, Cambridge University Pres (1982). With the kind pemission of the author and publishers.

表 7 皮尔逊相关系数的临界值

	单侧检验的显著性水平			
	0.05	0.025	0.005	0.0005
	双侧检验的显著性水平			
df (*N* − 2)	0.10	0.05	0.01	0.001
2	0.9000	0.9500	0.9900	0.9999
3	0.805	0.878	0.9587	0.9911
4	0.729	0.811	0.9172	0.9741
5	0.669	0.754	0.875	0.9509
6	0.621	0.707	0.834	0.9241
7	0.582	0.666	0.798	0.898
8	0.549	0.632	0.765	0.872
9	0.521	0.602	0.735	0.847
10	0.497	0.576	0.708	0.823
11	0.476	0.553	0.684	0.801
12	0.575	0.532	0.661	0.780
13	0.441	0.514	0.641	0.760
14	0.426	0.497	0.623	0.742
15	0.412	0.482	0.606	0.725
16	0.400	0.468	0.590	0.708
17	0.389	0.456	0.575	0.693
18	0.378	0.444	0.561	0.679
19	0.369	0.433	0.549	0.665
20	0.360	0.423	0.537	0.652
25	0.323	0.381	0.487	0.597
30	0.296	0.349	0.449	0.554
35	0.275	0.325	0.418	0.519
40	0.257	0.304	0.393	0.490
45	0.243	0.288	0.372	0.465
50	0.231	0.273	0.354	0.443
60	0.211	0.250	0.325	0.408
70	0.195	0.232	0.302	0.380
80	0.183	0.217	0.283	0.357
90	0.173	0.205	0.267	0.338
100	0.164	0.195	0.254	0.321

要达到各水平上的显著性水平，计算出的 r 值必须等于或超过表中的临界值。

SOURCE: F. C. Powell, *Cambridge Mathematical and Statistical Tables* (1976) Cambridge University Press. With kind permission of the author and publishers.

表 8 斯皮尔曼等级相关

	单侧检验的显著性水平			
	0.05	0.025	0.01	0.005
	双侧检验的显著性水平			
	0.10	0.05	0.02	0.01
$n=4$	1.000			
5	0.900	1.000	1.000	
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.755
12	0.503	0.587	0.671	0.727
13	0.484	0.560	0.648	0.703
14	0.464	0.538	0.622	0.675
15	0.443	0.521	0.604	0.654
16	0.429	0.503	0.582	0.635
17	0.414	0.485	0.566	0.615
18	0.401	0.472	0.550	0.600
19	0.391	0.460	0.535	0.584
20	0.380	0.447	0.520	0.570
21	0.370	0.435	0.508	0.556
22	0.361	0.425	0.496	0.544
23	0.353	0.415	0.486	0.532
24	0.344	0.406	0.476	0.521
25	0.337	0.398	0.466	0.511
26	0.331	0.390	0.457	0.501
27	0.324	0.382	0.448	0.491
28	0.317	0.375	0.440	0.483
29	0.312	0.368	0.433	0.475
30	0.306	0.362	0.425	0.467

如果 $n>30$,则对 ρ 的显著性检验可用下列公式:

$$t=\rho\sqrt{\frac{n-2}{1-\rho^2}} \qquad df=n-2$$

显著性水平查表 4

要达到各水平上的显著性水平,计算出的 ρ 值必须等于或大于表中的临界值。

SOURCE: J. H, Zhar, Significance testing of the Spearman Rank Correlation Coefficient, *Journal of the American Statistical Association*, 67, 578-580. Reprinted with permission from the *Journal of the American Statistical Association*.

参考文献

Abrahamsson, K. H., Berggren U., Hallberg-Lillemor, R. M. & Carlsson, S. G. (2002) Ambivalence in coping with dental fear and avoidance: a qualitative study. *Journal of Health Psychology*, 7, 6,653-64. 95

Ainsworth, M. D. S., Bell, S. M. & Stayton, D. J. (1971) Individual differences in strange situation behaviour of one-year-olds. In H. R. Schaffer (ed.) (1971) *The Origins of Human Social Relations*. London: Academic Press. 96

Anderson, C. (1987) Temperature and aggression: effects on quarterly, yearly, and city rates of violent and non-violent crime. *Journal of Personality and Social Psychology*, 52, 6, 1161-73. 35

Archer, J. (2000) Sex differences in aggression between heterosexual partners: a meta-analytic review. *Psychological Bulletin*, 126, 651-80. 11

Argyle, M. (1994) *The Psychology of Social Class*. London: Routledge. 185

Asch, S. E. (1956) Studies of independences and submission to group pressure. 1. A minority of one against a unanimous majority. *Psychological Monographs*, 70 (9) (Whole No. 416). 96

Awosunle, S. & Doyle, C. (2001) Same-race bias in the selection interview. *Selection Development Review*, 17, 3, 3-6. 86

Baker, L., Wagner, T. H. & Singer, S. (2003) Use of the Internet and e-mail for health care information: results from a national survey. *Journal of the American Medical Association*, 289, 18, 2400-406. 89

Bandura, A. (1965) Influence of models' reinforcement contingencies on the acquisition of imitative responses. *Journal of Personality and Social Psychology*, 1,589-95. 46, 62

Banyard, P. & Hunt, N. (2000) Reporting research: something missing? *The Psychologist*, 13, 2, 68-71. 6, 26

Barber, T. X. (1976) *Pitfalls in Human Research*. Oxford: Pergamon. 51

Benedict, R. (1934) *Patterns of Culture*. Boston: Houghton Mifflin. 59

Berry, J. W., Poortinga, Y. H., Segall, M. H. & Dasen, P. R. (2002) *Cross-Cultural Psychology: Research and Applications* (2nd edition). Cambridge: Cambridge University Press. 60

Bradley, D. R. (1991) Anatomy of a DATASIM simulation: the Doob and Gross horn-honking study. *Behavior Research Methods, Instruments and Computers*, 23, 2, 190-207. 204

Bramel, D. A. (1962) A dissonance theory approach to defensive projection. *Journal of Abnormal and Social Psychology*, 64, 121-9. 200

British Psychology Society (2000) *Code of conduct, ethical principles and guidelines*. Leicester: British Psychology Society. 196, 197-198, 202, 203, 205

Brody, G. H., Stoneman, Z. & Wheatley, P. (1984) Peer interaction in the presence and absence of observers. *Child Development*, 55, 1425-28. 66

Brown, R. (1988) *Group Processes: Dynamics Within and Between Groups*. Oxford: Blackwell. 46

Carlsmith, J., Elsworth, P. & Aronson, E. (1976) *Methods of Research in Social Psychology*. Reading, Mass.: Addison-Wesley. 48, 50

Charlesworth, R. & Hartup, W. W. (1967) Positive social reinforcement in the nursery school peer group. *Child Development*, 38, 993-1002. 66, 67

Cialdini, R. B., Reno, R. R. & Kallgren, C. A. (1990) A focus theory of normative conduct: recycling the concept of norms to reduce litter in public places. *Journal of Personality and Social Psychology*, 58, 1015-20. 54

Cook, T. D. & Campbell, T. T. (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally. 47, 55

Coolican, H. (2004) *Research Methods and Statistics in Psychology, 4th edition*. London: Hodder Arnold. 55, 87, 95, 119, 147, 176, 188, 197

Cumberbatch, G. (1990) *Television Advertising and Sex Role Stereotyping: A Content Analysis* (working paper Ⅳ for the Broadcasting Standards Council). Communications Research Group, Aston University. 99

Darley, J. M. & Latané, B. (1968) Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377-83. 10, 16

David, S. S. J., Chapman, A. J., Foot, H. C. & Sheehy, N. P. (1986) Peripheral vision and child pedestrian accidents. *British Journal of Psychology*, vol. 77, 4. 125

Doob, A. N. & Gross, A. E. (1968) Status of frustration as an inhibitor of horn-honking responses. *Journal of Social Psychology*, 76, 213-8. 204

Durrant, J. E. (2000) Trends in youth crime and well-being since the abolition of corporal punishment in Sweden. *Youth and Society*, 31, 4, 437-55. 191

Earley, P. C. (1989) Social loafing and collectivism: a comparison of the United States and the People's Republic of China. *Administrative Science Quarterly*, 34, 565-81. 46

Edwards, D. & Potter, J. (1992) *Discursive psychology*. London: Sage. 97

Ekéus, C., Christensson, K. & Hjern, A. (2004) Unintentional and violent injuries among pre-school children of teenage mothers in Sweden: a national cohort study. *Journal of Epidemiology and Community Health*, 58, 8, 680-85. 58

Elliott, R., Fischer, C. & Rennie, D. (1999) Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, 38, 215-29. 95

Eron, L. D., Huesmann, L. R. & Lefkowitz, M. M. & Walder, L. D. (1972) Does television violence cause aggression? *American Psychologist*, 27, 4, 253-63. 58

Felmet, M. B. (1998) The effects of karate training on the levels of attention and impulsivity of children with attention deficit/hyperactivity disorder. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 59, (4-A), 1077. 42

Gittelsohn, J., Shankar, A. V., West, K. P., Ram, R. M. & Gnywali, T. (1997) Estimating reactivity in direct observation studies of health behaviours. *Human Organization*, 56, 2, 182-89. 66

Giles, D. (2002) *Advanced Research Methods in Psychology*. Hove: Routledge. 95

Glaser, B. G. (1992) *Emergence vs Forcing: Basics of Grounded Theory Analysis*. Mill Valley CA: Sociology Press. 97

Glaser, B. G. & Strauss, A. L. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago IL: Aldine. 96

Goldberg, D. (1978) *The General Health Questionnaire*. Windsor: NFER-NELSON. 210, 212

Goldberg, P. (1968) Are women prejudiced against women? *Transaction*, April 1968. 228

Gregory, R. L. & Wallace, J. G. (1963) *Recovery from Early Blindness*. Cambridge: Heiffer. 100

Harré, R. (1981) The positivist-empiricist approach and its alternative. In R. Reason & J. Rowan (1981) *Human Enquiry: A Sourcebook of New Paradigm Research*. Chichester: Wiley. 93

Hayes, N. (1997) *Doing Qualitative Analysis in Psychology*. Hove: Psychology Press. 95, 98

Henwood, K. I. & Pidgeon, N. F. (1992) Qualitative research and psychological theorizing. *British Journal of Psychology*, 83, 97-111. 95

Hofling, C. K., Brotzman, E., Dalrymple, S., Graves, N. & Pierce, C. M. (1966) An experimental study in nurse-physician relationships. *Journal of Nervous and Mental Disease*, 143, 171-80.

Humphreys, L. (1970) *Tearoom Trade*. Chicago: Aldine. 48

Jansen, V. A. A., Stollenwerk, N., Jensen, H. J., Ramsay, M. E., Edmunds, W. J. & Rhodes, C. J. (2003) Measles outbreaks in a population with declining vaccine uptake. *Science*, 8, 301,804. 11

Johnston, W. M. & Davey, G. C. L. (1997) The psychological impact of negative TV news bulletins: the catastrophizing of personal worries. *British Journal of Psychology*, 88, 85-91. 201

Jolyon, L. (1962) *Hallucinations*. Oxford: Grune and Stratton. 202

Jones, F. & Fletcher, C. B. (1992) *Transmission of Occupational Stress: a Study of Daily Fluctuations in Work Stressors and Strains and their Impact on Marital Partners*. VIth European Health Psychology Society Conference (presented as poster), University of Leipzig (August). 94

Kagan, J., Kearsley, R. G. & Zelazo, P. R. (1980) *Infancy—Its Place in Human Development*. Cambridge, Mass.: Harvard University Press. 58

Kenrick, D. T. & MacFarlane, S. W. (1986) Ambient temperature and horn honking: a field study of the heat/aggression relationship. *Environment and Behavior*, 18, 2, 179-91. 40

Kimmel, A. J. (1998) In defence of deception. *American Psychologist*, 53, 7, 803-805. 200

Kinsey, A. C., Pomeroy, W. B., Martin, C. E. & Gebhard, P. H. (1953) *Sexual Behaviour in the Human Female*. Philadelphia: Saunders. 89

Klein, P. S. (1991) Improving the quality of parental interaction with very low birth weight children: a longitudinal study using a mediated learning experience model. *Infant Mental Health Journal*, 12, 4, 321-37. 204

Kvavilashvili, L. & Ellis, J. (2004) Ecological validity and real-life/laboratory controversy in memory research: a critical and historical review. *History of Philosophy and Psychology*, 6, 59-80. 48

Latané, B. & Darley, J. M. (1976) *Help in a Crisis: Bystander Response to an Emergency*. Morristown, NJ: General Learning Press. 200, 202

Lawn, S. J. (2004) Systemic barriers to quitting smoking among institutionalised public mental health service populations: a comparison of two Australian sites. *International Journal of Social Psychiatry*, 50, 3,204-15. 67, 95

Leyens, J., Camino, L., Parke, R. D. & Berkowitz, L. (1975) Effects of movie violence on aggression in a field setting as a function of group dominance and cohesion. *Journal of Personality and Social Psychology*, 32, 346-

60. 204

Likert, R. A. (1932) A technique for the measurement of attitudes. *Archives of Psychology*, 140, 55. 76-77, 273

Luria, A. R. (1969) *The Mind of a Mnemonist*. London: Jonathan Cape. 100

Maguire, E. A., Spiers, H. J. & Good, C. D. (2003) Navigation expertise and the human hippocampus: a structural brain imaging analysis. *Hippocampus*, 13, 2, 250-59. 3

McKillip, J. & Posavac, E. J. (1975) Judgments of responsibility for an accident. *Journal of Personality*, 43, 2,248-65. 211

Mead, G. H. (1934) *Mind, Self, and Society*. Chicago IL: University of Chicago Press. 93

Milgram, S. (1963) Behavioural study of obedience. *Journal of Abnormal and Social Psychology*, 63, 371-8. 198-203

Milgram, S. (1974) *Obedience to Authority*. New York: Harper and Row. 48, 96

Neisser, U. (1978) Memory: what are the important questions? In M. M. Gruneberg, P. E. Morris & R. N. Sykes (eds.) (1988) *Practical Aspects of Memory*. London: Academic Press. 46

Oliver, P. (2003) *The Student's Guide to Research Ethics*. Maidenhead: Open University. 206

Ora, J. P. (1965) Characteristics of the volunteer for psychological investigations. Office of Naval Research Contract 2149(03), Technical Report27. 31

Orne, M. T. (1962) On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776-83. 50

Ortmann, A. & Hertwig, R. (1997) Is deception acceptable? *American Psychologist*, 52, 7, 746-47. 200

Ortmann, A. & Hertwig, R. (1998) The question remains: is deception acceptable? *American Psychologist*, 53, 7, 806-807. 201

Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. (1957) *The Measurement of Meaning*. Urbana: University of Illinois. 77

Penfield, W. & Rasmussen, T. (1950) *The Cerebral Cortex of Man: a Clinical Study of Localization of Function*. New York: Hafner. 92

Perner, J., Leekam, S. R. & Wimmer, H. (1987) Three-year-olds' difficulty with false belief: the case for a conceptual deficit. *British Journal of Developmental Psychology*, 5, 2, 125-37. 58

Piaget, J. (1936) *The Origins of Intelligence in the Child*. London: Routledge. 87

Pichert, J. W. & Anderson, R. C. (1977) Taking different perspectives on a story. *Journal of Educational Psychology*, 69, 4, 309-15. 212

Piliavin, I. M., Rodin, J. & Piliavin, J. A. (1969) Good Samaritanism: an underground phenomenon? *Journal of Personality and Social Psychology*, 13,289-99. 204

Pulkki-Raback, L., Elovainio, M., Kivimaki, M., Raitakari, O. & Keltikangas-Jarvinen, L. (2005) Temperament in childhood predicts body mass in adulthood: the cardiovascular risk in young Finns study. *Health Psychology*, 24, 3, 307-15. 88

Rank, S. & Jacobson, C. (1977) Hospital nurses' compliance with medication overdose orders: a failure to replicate. *Journal of Health and Social Behaviour*, 18, 188-93. 48

Reason, P. & Rowan, J. (1981) (eds.) *Human Enquiry: A Sourcebook in New Paradigm Research*. Chichester: Wiley. 201,206

Reyner, L. A. &Horne, J. A (2002) Efficacy of a 'functional energy drink' in counteracting driver sleepiness. *Physiology and Behavior*, 75, 3, 331-35. 15

Richards, G. (1997) *'Race', Racism and Psychology*. London: Routledge.

Ring, K., Wallston, K. & Corey, M. (1970) Mode of debriefing as a factor affecting subjective reaction to a Milgram-type obedience experiment: an ethical inquiry. Representative Research in Social Psychology, 1, 67-88. 201

Ritov, I. & Baron, J. (1990) Reluctance to vaccinate: omission, bias and ambiguity. *Journal of Behavioral Decision Making*, 3, 263-77. 11

Robson, C. (2002) *Real World Research*. Oxford: Blackwell. 95

Roethlisberger, F. J. & Dickson, W. J. (1939) *Management and the Worker*. Oxford: Harvard University Press. 50

Rogers, C. R. (1961) *On Becoming a Person: a Therapist's View of Psychotherapy*. London: Constable. 97

Rorschach, H. (1921) *Psychodiagnostics: a Diagnostic Test Based on Perception*. Oxford: Grune and Stratton. 83

Rosenhan, D. L. (1973) On being sane in insane places. Science, 179, 250-8. 67, 92

Rosenthal, R. (1966) *Experimenter Effects in Behavioural Research*. New York: Appleton-Century-Crofts. 50

Rosenthal, R. & Jacobson, L. (1968) *Psychology in the classroom*. New York: Holt. 51

Rosnow, R. L. & Rosenthal, R. (1997) *People studying people: artifacts and ethics in behavioral research*. New York: W. H. Freeman. 31, 86

Ross, H. L., Campbell, D. T. & Glass, G. V. (1970) Determining the social effects of a legal reform: the British 'breathalyser' crackdown of 1967. *American Behavioral Scientist*, 13,493-509. 54

Rotter, J. B. (1966) Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 30, 1, 1-26. 84, 85, 210

Rule, B. G., Taylor, B. R. & Dobbs, R. A. (1987) Prim-

ing effects of heat on aggressive thoughts. *Social Cognition*, *5*, *2*, *131-43*. *38*, *39-40*, *41*, *46*, *53*

Russell, B. (1976) *The Impact of Science on Society*. London: Unwin Paperbacks. 8

Sears, D. (1986) College sophomores in the laboratory: influences of a narrow database on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515-30. 27

Shimizu, R. & Rutter, M. (2005) No effect of MMR withdrawal on the incidence of autism: a total population study. *Journal of Child Psychology and Psychiatry*, 46, 6, 572-79. 11

Shirav, E. & Levy, D. (2004) *Cross-cultural Psychology: Critical Thinking and Contemporary Applications*. Boston MA: Allyn and Bacon. 60

Shriberg, L. (1972) Intercorrelations among repression-sensitization, extroversion, neuroticism, social desirability, and locus of control. *Psychological Reports*, 31, 3, 925-26. 84

Sieber, J. E., Iannuzzo, R. & Rodriguez, B. (1995) Deception methods in psychology: have they changed in 23 years? *Ethics and Behavior*, 5, 1, 67-85. 200

Smith, J. A. (2003) *Qualitative Psychology: a Practical Guide to Research Methods*. London: Sage. 95, 96, 97

Smith, P. B. & Bond, M. H. (2005) *Understanding Social Psychology Across Cultures: Living and Working in a Changing World*. London: Sage. 60

Smith, P. B., Trompenaars, F. & Dugan, S. (1995) The Rotter locus of control scale in 43 countries: a test of cultural relativity. *International Journal of Psychology*, 30, 377-400. 85

Smyth, T. R. (2004) *The Principles of Writing in Psychology*. Palgrave Macmillan. 216

Stipek, D. (1998) Differences between American and Chinese in the circumstances evoking pride, shame and guilt. *Journal of Cross-Cultural Psychology*, 29, 5, 616-29. 60

Strauss, A. L. & Corbin, J. A. (1990) *Basics of qualitative research; grounded theory procedures and techniques*. Newbury Park CA: Sage. 97

Taylor, K. M. & Shepperd, J. A. (1996) Probing suspicion among participants in deception research. *American Psychologist*, 51, 8, 886-87. 201

Thigpen, C. H. & Cleckley, H. (1954) A case of multiple personality. *Journal of Abnormal and Social Psychology*, 49, 175-81. 100

Torbert, W. R. (1981) Why educational research has been so uneducational: the case for a new model of social science based on collaborative enquiry. In P. Reason & J. Rowan (eds.) (1981) *Human Enquiry: A Sourcebook in New Paradigm Research*. Chichester: Wiley. 203

Triplett, N. (1898) The dynamogenic factors in pacemaking and competition. *American Journal of Psychology*, 9, 505-23. 135

Walster, E. (1966) Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, 3, 1, 73-79. 211

Wardle, J. & Watters, R. (2004) Sociocultural influences on attitudes to weight and eating: results of a natural experiment. *International Journal of Eating Disorders*, 35, 4, 589-96. 54-55

Watson, J. B. & Rayner, R. (1920) Conditioned emotional reactions. *Journal of Experimental Psychology*, 3, 1-14. 92, 202

White, W. F. (1943) *Street Corner Society: the Social Structure of an Italian Slum*. Chicago: The University of Chicago Press. 67

Wigal, J. K., Stout, C. & Kotses, H. (1997) Experimenter expectancy in resistance to respiratory air flow. *Psychosomatic Medicine*, 59, 3, 318-22. 51

Willig, C. (2001) *Introducing Qualitative Research in Psychology*. Buckingham: Open University Press. 95

Wilson, S. R., Brown, N. L., Mejia, C. & Lavori, P. (2002) Effects of interviewer characteristics on reported sexual behavior of California Latino couples. *Hispanic Journal of Behavioral Sciences*, 24, 1, 38-62. 86

Wimmer, H. G. S. & Perner, J. (1985) Young children's conception of lying: moral intuition and the denotation and connotation of 'to lie'. *Developmental Psychology*, 21, 6, 993-95. 59

Yardley, L. (2000) Dilemmas in qualitative health research. *Psychology and Health*, 15, 215-28. 95

Zanna, M. P. & Cooper, J. (1974) Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703-709. 86

Zimbardo, P. G. (1972) Pathology of imprisonment. *Society*, April 1972. 63, 202

万卷方法总书目

万卷方法是我国第一套系统介绍社会科学研究方法的大型丛书，来自中国社科院、北京大学等研究机构和高校的两百余名学者参与了丛书的写作和翻译工作。至今已出版图书85个品种，其中绝大多数是2008年以来出版的新书。

85 社会科学方法论（国家十二五规划教材）
978-7-5624-6204-0

84 田野工作的艺术
978-7-5624-6257-6

83 图解AMOS在学术研究中的应用
978-7-5624-6223-1

82 应用STATA做统计分析（更新至STATA10.0）
978-7-5624-4483-1

81 社会调查设计与数据分析——从立题到发表
978-7-5624-6074-9

80 质性研究导引
978-7-5624-6132-6

79 APA格式——国际社会科学学术写作规范手册
978-7-5624-6105-0

78 如何做心理学实验
978-7-5624-6151-7

77 话语分析导论
978-7-5624-6075-6

76 心理学学位论文写作全程指导
978-7-5624-6113-5

75 心理学研究方法导论
978-7-5624-5828-9

74 分类数据分析
978-7-5624-6133-3

73 结构方程模型：AMOS的操作与应用（附光盘版）
978-7-5624-5720-6

72 AMOS与研究方法（第2版）
978-7-5624-5569-1

71 爱上统计学（第2版）
978-7-5624-5891-3

70 社会科学定量研究的变量类型、方法选择与范例解析
978-7-5624-5714-5

69 案例研究：设计与方法（中译第2版）
978-7-5624-5732-9

68 问卷设计手册：市场研究、民意调查、社会调查、健康调查指南
978-7-5624-5597-4

67 广义潜变量模型：多层次、纵贯性以及结构方程模型
978-7-5624-5393-2

66 调查问卷的设计与评估
978-7-5624-5153-2

65 心理学论文写作——基于APA格式的指导
978-7-5624-5354-3

64 心理学质性资料的分析
978-7-5624-5363-5

63 问卷统计分析实务：SPSS操作与应用
978-7-5624-5088-7

62 如何做综述性研究
978-7-5624-5375-8

61 质性访谈方法
978-7-5624-5307-9

60 量表编制：理论与应用（校订新译本）
978-7-5624-5285-0

59 质性研究：反思与评论（第2卷）
978-7-5624-5143-3

58 实验设计原理：社会科学理论验证的一种路径
978-7-5624-5187-7

57 混合方法论：定性研究与定量研究的结合
978-7-5624-5110-5

56 社会统计学
978-7-5624-5253-9

55 校长办公室的那个人（质性研究个案阅读）
978-7-5624-4880-8

54 泰利的街角（质性研究个案阅读）
978-7-5624-4937-9

53 客厅即工厂（质性研究个案阅读）
978-7-5624-4886-0

52 标准化调查访问
978-7-5624-5062-7

51 解释互动论
978-7-5624-4936-2

50 如何撰写研究计划书
978-7-5624-5087-0

49 质性研究的理论视角：一种反身性的方法论
978-7-5624-4889-1

48 社会评估：过程、方法与技术
978-7-5624-4975-1

47 如何解读统计图表
978-7-5624-4906-5

46 公共管理定量分析：方法与技术（第2版）
978-7-5624-3640-9

45 量化研究与统计方法
978-7-5624-4821-1

44 心理学研究要义
978-7-5624-5098-6

43 调查研究方法（校订新译本）
978-7-5624-3289-0

42 分析社会情境:质性观察和分析方法
978-7-5624-4690-3

41 建构扎根理论:质性研究实践指南
978-7-5624-4747-4

40 参与观察法
978-7-5624-4616-3

39 文化研究:民族志方法与生活文化
978-7-5624-4698-9

38 质性研究方法:健康及相关专业研究指南
978-7-5624-4720-7

37 如何做质性研究
978-7-5624-4697-2

36 质性研究中的访谈:教育及社会科学研究者指南
978-7-5624-4679-8

35 案例研究方法的应用(中译第2版)
978-7-5624-3278-3

34 教育研究方法论探索
978-7-5624-4649-1

33 实用抽样方法
978-7-5624-4487-9

32 质性研究:反思与评论(第1卷)
978-7-5624-4462-6

31 社会科学研究的思维要素(第8版)
978-7-5624-4465-7

30 哲学史方法论十四讲
978-7-5624-4446-6

29 社会研究方法
978-7-5624-4456-5

28 质性资料的分析:方法与实践(第2版)
978-7-5624-4426-8

27 实用数据再分析法(第2版)
978-7-5624-4296-7

26 质性研究的伦理
978-7-5624-4304-9

25 叙事研究:阅读、倾听与理解
978-7-5624-4303-2

24 质化方法在教育研究中的应用(第2版)
978-7-5624-4349-0

23 复杂调查设计与分析的实用方法(第2版)
978-7-5624-4290-5

22 研究设计与写作指导:定性、定量与混合研究的路径
978-7-5624-3644-7

21 做自然主义研究:方法指南
978-7-5624-4259-2

20 多层次模型分析导论(第2版)
978-7-5624-4060-4

19 评估:方法与技术(第7版)
978-7-5624-3994-3

18 焦点团体:应用研究实践指南(第3版)
978-7-5624-3990-5

17 质的研究的设计:一种互动的取向（第2版）
978-7-5624-3971-4

16 组织诊断:方法、模型和过程(第3版)
978-7-5624-3055-1

15 民族志:步步深入(第2版)
978-7-5624-3996-7

14 分组比较的统计分析(第2版)
978-7-5624-3942-4

13 抽样调查设计导论（第2版）
978-7-5624-3943-1

12 定性研究(第3卷):经验资料收集与分析的方法(2版)
978-7-5624-3944-8

11 定性研究(第4卷):解释、评估与描述(第2版)
978-7-5624-3948-6

10 定性研究(第1卷):方法论基础(第2版)
978-7-5624-3851-9

9 定性研究(第2卷):策略与艺术（第2版）
978-7-5624-3286-9

8 社会网络分析法（第2版）
978-7-5624-2147-4

7 公共政策内容分析方法:
978-7-5624-3850-2

6 复杂性科学的方法论研究
978-7-5624-3825-0

5 社会科学研究:方法评论
978-7-5624-3689-8

4 论教育科学:基于文化哲学的批判与建构
978-7-5624-3641-6

3 科学决策方法:从社会科学研究到政策分析
7-5624-3669-0

2 电话调查方法:抽样、筛选与监控（第2版）
7-5624-3441-7

1 研究设计与社会测量导引(第6版)
978-7-5624-3295-1

为了建设好“万卷方法”，更好地服务学界，现由重庆大学出版社和人大经济论坛做出决定，凡购买重庆大学出版社的万卷方法系列图书的读者，填写以下信息调查表（复印即可），邮寄给我们（400030　重庆大学出版社　林佳木），经过认证后，我们将会赠送人大经济论坛币100个（可免费下载丛书相关学习资料并与教师及学友进行交流）：

读者情况调查表	
姓名	
单位	
联系电话	
E-mail	
论坛 ID	
使用书籍	
购买渠道	
对丛书建设的建议	
邮政地址（邮编）	

人大经济论坛

——国内最大的经济、管理、金融、统计类在线教育网站

人大经济论坛（网址：http://www. pinggu. org）依托中国人民大学经济学院，于2003年成立，致力于推动经济学科的进步，传播优秀教育资源，目前已经发展成为国内最大的经济、管理、金融、统计类的在线教育和咨询网站，也是国内最活跃和最具影响力的经济类网站。

1. 拥有国内经济类教育网站最多的关注人数，注册用户以百万计，日均数十万经济相关人士访问本站。

2. 是国内最丰富的经管类教育资源共享数据库和发布平台。

3. 论坛给所有会员提供学术交流与讨论的平台，同时也有网络社交 SNS 的空间，经管百科提供了丰富专业的经管类在线词典，数据定制和数据处理分析服务是您做实证研究的好帮手，免费的经济金融数据库使您不再为数据发愁，更有完善的经管统计类培训和教学相关软件，只要您是学习、研究或从事经管类行业，人大经济论坛就能满足您的需要！